Modelling pathological processes from heterogeneous and high-dimensional biomedical data

Marco Lorenzi

UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS
École Doctorale STIC

Mémoire présenté pour l'obtention de
L'HABILITATION À DIRIGER DES RECHERCHES

Marco Lorenzi

Modelling pathological processes from heterogeneous and high-dimensional biomedical data

Jury Members:

Prof.	John	**Ashburner**	University College London	*Rapporteur*
Prof.	Olivier	**Colliot**	Centre national de la recherche scientifique	*Rapporteur*
Prof.	Daniel	**Rueckert**	Imperial College London	*Rapporteur*
Prof.	Giovanni B.	**Frisoni**	Geneva University Hospitals	*Examinateur*
Prof.	Sebastien	**Ourselin**	King's College London	*Examinateur*
Prof.	Xavier	**Pennec**	Inria	*Examinateur*
Prof.	Nicholas	**Ayache**	Inria	*President*

Defense date: **January 10th, 2020**

I would first like to thank the members of my jury, Professors Ashburner, Colliot and Rueckert. It has been a great honor to defend my HDR in front of such a unique panel of exemplary researchers. Since my earliest steps in research, their activity has continuously guided me and inspired my work. Being able to present my research project in front of them is a unique opportunity which became reality.

I would like to give a special thanks to the advisors that followed me through the years, Professors Frisoni, Pennec, Ayache and Ourselin. I owe to them most of my research achievements, and I am deeply grateful to them for the great support they always gave to me, from both professional and human perspectives. I have learned a lot from them, and I still do now. A particular thank you goes to all the current and past colleagues from the Asclepios/Epione team, with whom I've been sharing my working life in a very enriching and special environment, and to my colleagues at TIG and CMIC in UCL, and at FBF. There are so many beautiful memories and stories that I had the luck to collect throughout the years, all thanks to the great people I had the chance to meet during my journey.

Finally, I would like to thank my family, my mum and my dad, who always supported me during my erratic path, and my brother. A last special thanks is for Maria, who helps me everyday to become a better person.

Marco

Contents

Chapter I

Introduction

Neurodegenerative diseases (NDD) are a broad class of brain disorders leading to molecular, functional, and structural brain damage, ultimately causing steady cognitive decline and drastic impairment of daily living activities. There are currently over 50 million people worldwide affected by NDD, such as Alzheimer's disease (AD), for an overall cost of more than US $818 billion per year [The17]. This number is expected to triple by 2050 and there are no effective treatments that can stop the spread of damage in the brain. NDD research targets a global societal and economical challenge. The urgency of this problem is exemplified by the decision on December 2015 of the American government to allocate the record sum of 1 billion USD for research and cure of Alzheimer's disease.

Clinical evidence suggests that early treatment of NDD is necessary to lead to effective disease modifying pharmacological strategies. However, early stages of NDD are usually characterized by subtle individual changes and high variability across individuals, making it extremely challenging to identify early pathological traits for better treatment and testing of novel drugs. Clinicians are asking for precise models defining optimal sets of measurements (and combinations of them) to uniquely identify pathological traits in patients. Pharmaceutical companies need instead to better identify clinical populations for testing disease modifying drugs, as well as to identify novel targets for the development of pharmacological treatments.

I.1 NDD modeling for improving care and treatment

The clear understanding of neurodegeneration is currently beyond our reach. Neurodegenerative processes, such as AD, are characterized by interrelated pathological alterations of the brain's structure, function and biology. Our understanding of NDD is therefore tied to the ability of linking knowledge from domains today still largely considered separately, from cellular and molecular biology to neuroscience.

Current lack of satisfactory biological models of neurodegeneration frustrates our advance-

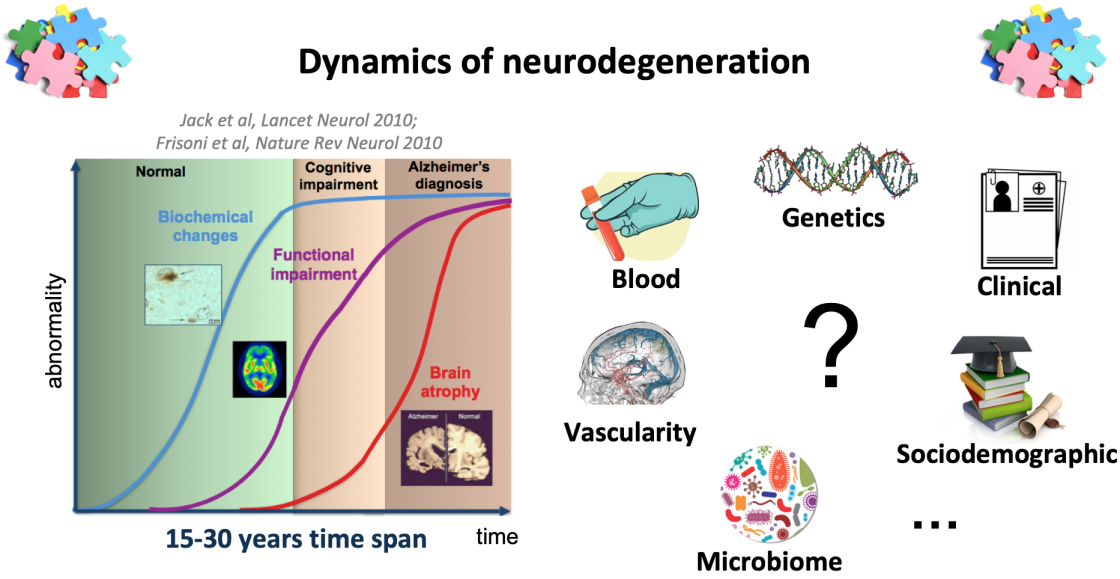


Figure I.1: Biomedical data analysis in NDD research: heterogeneity and complexity.

ments in the understanding and treatment of NDD. To compensate for this, scientists are currently gathering insights on the pathological mechanisms underlying NDD through the interrogation of large-scale and heterogeneous biomedical dataset. The field of NDD research is currently driving ambitious data collection initiatives aimed at providing scientists with data samples representing the full spectrum of the pathological process: brain imaging data, genetic, blood, life-style, and clinical assessment. As a result, computer science is currently playing a central role in NDD research through the development of quantitative computational tools leveraging on currently available biobank and clinical studies. This is exemplified by the large amount of literature published in the last years focusing on the study of brain images and clinical inormation from patients and healthy controls available from international databases, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [PAB+10], or the UK Biobank[SGA+15].

Today we enjoy a unique opportunity to study the relationship between clinical conditions and brain changes on data of unprecedented dimension. Based on this unique amount of available clinical and imaging data, the clinical questions we want to answer are the following. (i) Can we accurately quantify when, how, and where NDD manifest in the brain? (ii) Can we predict the course of the pathology in a given individual, along with the expected effect of disease modifying drugs? Answering these questions would ultimately require to develop a computational model simulating the joint temporal evolution of interdependent biological processes, while accounting for the disparate variability observable in the general population, from genetics to life-style factors.

I.2 The challenge of NDD modeling

The heterogeneity and complexity of biomedical data is a crucial pitfall for modeling neurodegeneration in clinical studies. In NDD research we tackle the challenge of modeling datasets composed by collections of brain imaging, biological samples, and clinical scores/socio-demographic information (Figure I.1). To give an example of the sheer volume of data these studies entail, the storing and analysis of this amount of currently available information from the UK Biobank requires several terabytes of disk space on dedicated servers, along with tremendous computational power for processing [SN18].

Imaging data

High-resolution 3D brain images have dimensionality in the order of millions of voxels structured in 3D grids. Imaging data may represent a variety of anatomical and molecular information, such as brain gray/white matter integrity [AF00b] from magnetic resonance imaging (MRI), and metabolism, or spatial distribution of pathological proteins from positron emission tomography (PET). Each modality is characterized by specific signal resolution and intensity properties, that must be accounted for by means of specific modeling hypothesis and constraints [HMB+16, YHG+17]. Images are characterized by geometric variability of brain shape and size, and by signal correlation patterns varying according to disease stage and brain locations. Correlation properties depend either on the local information on neighborhoods around voxels, or on non-local relationships across separated, but functionally interdependent, brain regions.

Time series of brain images allow investigating subtle pathological evolutions of the brain, and introduce an additional temporal dimension leading to important methodological challenges. First, in prospective clinical studies only a handful of serial measures are typically available for each patient. Second, the disease timing and speed is highly heterogeneous across individuals [YFL+11]. This implies that patients' time series cannot be generally compared with respect to an absolute time reference, such as the chronological age. The temporal analysis thus further requires to account for individual changes in time scale [SACD15]. For all these reasons, standard tools for time-series analysis developed in statistics cannot find straightforward application in this context.

Biological information

Biological information represents a wide range of processes and mechanisms, related for instance to proteins or lipids concentrations in the blood, or to genomic variation. This information is usually encoded in high dimensional arrays (up to millions of features), with non-trivial and mostly unknown correlation properties. For example, single nucelotide polymorhisms (SNP) are variants of a single nucleotide basis in a specific genomic position that occour with a certain frequency in the general population (e.g. $> 1\%$) [Nac01, C+15]. There are currently hundreds

of millions of known SNPs. A typical question is whether the presence of a specific SNP, or of a combination of them, is associated to the prevalence of a pathological trait (called phenotype). To tackle this problem, researchers aim at jointly studying millions of SNP counts measured in a population to define a *genetic signature* of the disease phenotype [Dud13, TWT18, CSGC16]. SNPs can be therefore analyzed under different assumptions about their inter-dependency, for instance concerning their proximity on the genetic code, or their joint involvement in known biological processes, called *pathways* [BWSL04, SJH+12]. The complexity of this kind of analyses is exacerbated when trying to merge these different sources of information into a joint model, as it is done for example in the field of *imaging-genetics*. In all these case, large-scale interdependencies occurring within and across modalities must be identified through the analysis of typically very high-dimensional data, oftentimes in the moderate to low-sample size regime characterizing clinical datasets.

The statistical treatment of heterogeneity and high-dimensions in biomedical studies is a challenging problem. Small simple size characterizing clinical studies often prevents the application of typical "data-hungry" machine learning methods, such as deep neural networks. To cope with this issue, tractability is usually achieved by narrowing down the modeling problem to the analysis of partial aspects of the disease. This is done either by modeling a limited amount of information (e.g. brain gray matter changes only), or by considering narrow time windows of the disease history (e.g. data for specific diagnostic groups). Another common assumption consists in making simplistic and conservative assumptions about the biomedical data properties. Mass independence between features is a widely adopted hypothesis [SN18]. While leading to simplified implementation schemes, this kind of assumption entails delicate statistical consequences, such as the multiple comparison problem, ultimately leading to sensitivity loss, and lack of generalization and interpretability [Dav04].

It is an exciting time for a data scientist to work in the field of biomedical data analysis. Reliably modeling heterogeneous and large scale biomedical signals represents a fundamental step towards the understanding of pathological processes and disease dynamics in populations. Achieving this goal requires innovative strategies for modeling data of unprecedented complexity and dimension, and involves challenges at the crossing of several scientific disciplines. It is within this stimulating environment that I have had the chance to develop my research project.

I.3 My research path

My research focuses on the analysis of spatio-temporal multimodal medical images and, more generally, on the development of methods to jointly analyze heterogenous biomedical information. In particular, I have been designing and implementing frameworks based on computational anatomy and statistical learning, to model and provide quantitative information about subtle neurological changes. My aim is twofold: to provide statistical models of the disease trajectory in dementia across the whole time course of NDD; and to develop techniques enabling the

exploitation of heterogeneous biomedical signal, which could provide a more complete picture of the pathology.

To this end, my contributions are in the domains of statistical learning, mathematics, and neuroscience, with a special focus on translation. An important part of my publications is based on collaborations with clinical partners such as Dr. Giovanni B. Frisoni (University of Geneva, CH), the Centre for Medical Image Computing (CMIC) of University College London (UCL, London, UK), and the Memory Center of Nice (France).

The ensemble of my research can be broadly assimilated to three main research axes: *Modeling and quantifying temporal changes in biomedical data* (Axis I), *Handling heterogeneity in biomedical data* (Axis II), and *Federated statistical learning for meta-analyses of biomedical data* (Axis III). This manuscript represents a digest of these research directions, and in what follows each axes is introduced and discussed in a dedicated chapter. The related bibliography is provided in Appendix.

I.4 Research axis summary

Axis I. Modeling and quantifying temporal changes in biomedical data

This axis addresses the problem of modeling time-series of biomedical measures in clinical data. Differently from classical time-series analysis problems, the repeated measurements we wish to model are usually of very high-dimension, and characterized by complex correlation properties. This problem requires the development of statistical approaches able to scale to spatio-temporally correlated high-dimensional signals, such as medical images. Moreover, longitudinal clinical NDD dataset generally lack of a well-defined temporal reference, since the onset of the pathology may vary across individuals according to genetic and environmental factors. Therefore, age or visit date information are biased time references for the individual longitudinal measurements. There is a critical need to define the AD evolution in a data-driven manner with respect to an absolute time scale associated to the natural history of the pathology.

Selected bibliography

- Modeling and inference of spatio-temporal protein dynamics across brain networks. Sara Garbarino and Marco Lorenzi. *Information Processing in Medical Imaging (IPMI)*, 57-69, 2019.

- Monotonic Gaussian Process for Spatio-Temporal Disease Progression Modeling in Brain Imaging Data. Clément Abi Nader, Nicholas Ayache, Philippe Robert, Marco Lorenzi. *NeuroImage*, 2019, to appear.

- Constraining the Dynamics of Deep Probabilistic Models. Marco Lorenzi and Maurizio Filippone. *Proceedings of the 35th International Conference on Machine Learning* (ICML),

PMLR 80:3233-3242, 2018.

- Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander and Sebastien Ourselin. *NeuroImage*, S1053-8119(17)30706-1, 2017.

- Efficient Gaussian process Based Model of Spatio-Temporal Changes in Time Series of Images. Marco Lorenzi, Gabriel Ziegler, Daniel Alexander and Sebastien Ourselin. *Information Processing in Medical Imaging (IPMI)*, 24:626-37, Springer, LNCS, 2015.

Axis II. Handling heterogeneity in biomedical data

This axis is related to the general problem of consistently integrating and analysing heterogeneous data sources. Clinical diagnosis in NDD is still mostly established through the visual grading of brain images, along with the assessment of clinical and biological data [DFJ+14]. To support clinicians in the analysis and interpretation of complex data sources, we are asked to develop computational methods for the identification and quantification of significant relationships across this heterogeneous biomedical information. Analysing this kind of heterogeneity involves important statistical challenges, related to uncertainty quantification, interpretability, and generalization of the results obtained with multivariate learning methods.

Selected bibliography

- Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data. Luigi Antelmi, Nicholas Ayache, Philippe Robert, Marco Lorenzi. *Proceedings of the 36th International Conference on Machine Learning*, ICML, 2019.

- Large Scale Cardiovascular Model Personalisation for Mechanistic Analysis of Heart and Brain Interactions. Jaume Banus Cobo, Marco Lorenzi, Oscar Camara, Maxime Sermesant. *Functional Imaging and Modeling of the Heart*, (FIMH), 2019.

- Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease: Evidence from functional prioritization in imaging genetics. Marco Lorenzi, Andre Altmann, Boris Gutman, Selina Wray, Charles Arber, Derrek P. Hibar, Neda Jahanshad, Jonathan M. Schott, Daniel C. Alexander, Paul M. Thompson and Sebastien Ourselin. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), 115(12):3162-3167, 2018.

- Multimodal Image Analysis in Alzheimer's Disease via Statistical Modelling of Non-local Intensity Correlations. Marco Lorenzi , Ivor J. Simpson, Alex F. Mendelson, Sjoerd B. Vos, M. Jorge Cardoso, Marc Modat, Jonathan Schott, Sebastien Ourselin. *Scientific Reports*, 6, 2261, 2016.

Axis III. Federated statistical learning for meta-analyses of biomedical data

This axis aims at exploring the capabilities of federated learning for deploying statistical learning methods on large-scale multi-centric biomedical data. My goal is to exploit the power of modern Bayesian learning methods at full capacity within the current clinical data scenario. The ambition of this axis requires methodological, technical, and translational advances.

This research area is gaining momentum thanks to recent initiatives from the large tech companies, calling growing attention from the media. For example, Google is currently experimenting federated learning for improving keyboard interactions in Android phones. Despite the recent industrial focus, this methodology is at its earliest applications, and still comes with important challenges, concerning communication bottlenecks, complexity of the training schemes, and high-computational demand for the end users.

The potential of this novel research path is witnessed by the upcoming projects recently funded through public initiatives (ANR project Fed-BioMED, EU funded PhD fellowship BoosUr-Career), and through industry collaborations (PhD project funded by Accenture Labs)[1].

Selected bibliography

- Multivariate Learning in Distributed Biomedical Databases: Meta-analysis of Large-scale Brain Imaging Data. Santiago Silva, Boris Gutman, Barbara Bardoni, Paul M Thompson, Andre Altmann, Marco Lorenzi. *IEEE International Symposium on Biomedical Imaging* (ISBI), Venice, 2019

- Secure multivariate large-scale multi-centric analysis through on-line learning. Marco Lorenzi, Boris Gutman, Paul M Thompson, Daniel C Alexander, Sebastien Ourselin and Andre Altmann. *12th International Symposium on Medical Information Processing and Analysis*, 1016016, International Society for Optics and Photonics, 2017.

[1] https://www.accenture.com/fr-fr/accenture-lab-sophia-antipolis

I.5 Supervision, Responsibilities & Other Research Activities

Current supervision

PhD student supervision

- Luigi Antelmi. Starting date 01/2018, 80% supervision, with Prof. N. Ayache, Inria. The PhD project of Luigi Antelmi is within the MNC3 initiative funded by Université Côte d'Azur (PI Prof. Philippe Robert and Prof. Nicholas Ayache). With Luigi I am extending my current work on the study of methods allowing to model the complex relationship between high-dimensional data, along with the associated uncertainty.

- Clément Abi Nader. PhD Starting date: 01/2018, 80% supervision, with Prof. N. Ayache, Inria. The PhD project of Clément Abi Nader is within the MNC3 initiative funded by the University Côte d'Azur (PI Prof. Philippe Robert and Prof. Nicholas Ayache). This project leverages on my expertise in spatio-temporal modeling of brain changes for the differential analysis of healthy aging and pathological brain processes in large-scales imaging databases.

- Jaume Banus Cobo. Starting date 11/2017, 50% supervision (with Dr. Maxime Sermesant, Inria). The PhD of Jaume is a multidisciplinary project for the joint analysis of brain and cardiac imaging data, funded by the University of Nice-Sophia Antipolis. The project is in collaboration with Prof. Oscar Camara (Universitat Pompeu Fabra, Spain). It aims at coupling machine learning and biophysical modeling methods for studying the joint relationship between cardiovascular abnormality and brain damage.

- Giorgos Lazaridis. Starting date 11/2017, I am supervising the methodological developments under the guidance of Prof. MD. D. Garway-Heath, UCL. This project is the result of my long-lasting collaboration with UCL. In 2015 Prof. D. Garway-Heath (UCL Institute of Ophthalmology) proposed me the co-supervision of a PhD project on the analysis of clinical trials data in glaucoma therapies. The PhD is funded by the pharmaceutical company Santen Pharmaceutical Co. Giorgos is applying machine learning methods for super-resolution and multimodal image analysis in the context of biomarker development in glaucoma and eye therapies.

- Raphaël Sivera. Starting date 04/2015 – Defense date 11/2019, co-supervision with Prof. N. Ayache and Prof. H. Delingette (supervisors) and Prof. X. Pennec (co-supervisor).

The work of Raphaël consists in the development of image-based biomarkers of the structural changes in Alzheimer's disease, from the analysis of collections of longitudinal MRI data. This project is partly based on modeling tools and theory developed during my PhD at Inria (Asclepios research group).

Post-doctoral fellow supervision

- Dr. Sara Garbarino. Starting date 05/2018 (2 years UCA funded postdoctoral fellowship). The project of Sara consists in the development of novel techniques for the modeling of the progression of neurodegenerative disorders through the study of propagation dynamics of pathological proteins across brain networks and pathways. In particular, we tackle current limitations of data-driven modeling methods by constraining the model dynamics through the introduction of biologically inspired constraints. The work is a natural extension of my latest contributions in constrained Gaussian process modeling.

Master student supervision

- Santiago Smith Silva Rincon. 01/05/2018 - 01/09/2018. Santiago was enrolled for a 6 months internship within the UCA funded project Meta-ImaGen. His internship consisted in developing a software platform for distributed multivariate analysis of brain imaging data. His project allowed me to propose larger scale scientific initiatives, recently funded by public agencies (Agence nationale de la recherche, ANR[2], and EU H2020 programme boostUrCAreer)[3] and by Industrial partners (Accenture Labs[4]).

Future supervision

Phd Student

- Yann Fraboni. Expected starting date 03/2020. Yann will be working on the analysis of bias in distributed machine learning applications. This project is funded by Accenture Labs of Sophia Antipolis.

- Etrit Haxholli. Expected starting date 03/2020. Etrit will be working on the problem of time-series analysis of graph supported data. This project is funded by the company MyDataModels[5] located in Sophia Antipolis.

[2] https://anr.fr
[3] http://univ-cotedazur.fr/fr/recherche/boosturcareer/
[4] https://www.accenture.com/fr-fr/accenture-lab-sophia-antipolis
[5] https://www.mydatamodels.com

- Santiago Smith Silva Rincon. Expected starting date 01/2020. Santiago will be working on the extension of his previous internship project, concerning the deploying of distributed statistical learning in multi-centric clinical studies. The project is funded by the EU H2020 programme boostUrCAreer[6].

Postdoctoral Fellow

- The ANR project Fed-BioMed will allow me to recruit a postdoctoral fellow for working on the development of federated Bayesian learning methods for the application in multi-centric clinical studies.

Past co-supervision

- Sebastiano Ferraris, PhD defended in 2018 within the Translational Imaging Unit (TIG), of University College London. Supervisor: Prof. Tom Vercauteren. The project concerned the development of non-linear registration methods for real-time motion tracking of unborn babies with severe birth defects.

- Mehdi Adj-Hamou, PhD defended in December 2016 within the Asclepios Team, Inria Sophia Antipolis. Supervisor: Prof. Nicholas Ayache. The PhD project concerned the modeling of the brain anatomical changes in adolescence and childhood, and extended the theory and algorithms developed during my PhD.

- Bishesh Khanal, PhD defended in July 2016 within the Asclepios Team, Inria Sophia Antipolis. Supervisor: Prof. Xavier Pennec. The PhD project concerned the modeling and simulation of realistic longitudinal atrophy patterns due to Alzheimer's disease in structural brain MRIs.

I.5.1 Collaborative Projects and Funding

- I am principal investigator of the project Fed-BioMed, that will be funded from 2020 to 2024 by French National Research Agency (200'000€). The project will focus on the development of federated learning methods for the analysis of private and large-scale multi-centric biomedical data.

- I was awarded in 2019 of a chair at the newly funded French Interdisciplinary Institute for Artificial Intelligence[7] (3IA).

[6] http://univ-cotedazur.fr/fr/recherche/boosturcareer/

[7] http://univ-cotedazur.fr/contenus-riches/actualites/en/universite-cote-dazur-selected-to-set-up-an-interdisciplinary-institute-for-artificial-intelligence

- I am scientific collaborator of the startup MyDataModels. In 2018 and 2019 I was scientific consultant providing expertise in the benchmarking and development of their technology. From 2020 the company will fund a PhD project under my supervision.

- I am scientific collaborator of the tech company Accenture Labs. From 2020 the company will fund a PhD project under my supervision.

- I am principal investigator of the international project Meta-ImaGen, for the distributed analysis of imaging-genetics data in meta studies. The project is funded by Université Côte d'Azur (38'700€) and fosters the collaboration between Inria, IPMC (Prof. Barbara Bardoni), University of Southern California (Prof. P. Thompson & Dr. B. Gutman), and University College London (Dr. A. Altmann).

- I was principal investigator of the project Big Data for Brain Research, funded in 2017 by the Department des Alpes Maritimes (AAP Santé 2017). The project aimed at creating a computing platform within the facility of Inria Sophia Antipolis dedicated to the analysis of large biomedical datasets. The funding was of 37'514€ and allowed the acquisition of server storage, system memory, and computing facilities (GPU cluster). I coordinated the realization of the data management system and computational platform.

- I am collaborator within the UCA funded project MNC3, through the co-supervision of 2 PhD students. I also contribute to the methodological workpackages of the H2020 international project EUROPOND (PI Prof. Daniel C. Alexander, UCL), and of the H2020-EFPIA funded project AMYPAD (PI Prof. Frederik Barkhof, VU University Medical Centre & UCL)

- I obtained funding (12'000GBP) and organized the 1st International Workshop on Modeling the Progression Of Neurological Diseases (POND2016), held in UCL in 2016. The 2 days invitation-only event hosted prominent scientists working on spatio-temporal modeling of neurodegenerative diseases (50 participants in total).

I.5.2 Scientific Engagement

- I am currently teaching at the Data Science Master[8] of Université Côte d'Azur (M1, 2nd semester: Model Selection and Resampling Methods (30 hours); M2, 1st semester: Bayesian Learning (30 hours)). The courses material is fully avilable online[9].

- I member of the scientific board of the UCA NeuroMod Institute for Modeling in Neuroscience and Cognition[10]. The institute is composed by more than 150 researchers and

[8] http://univ-cotedazur.fr/en/idex/formations-idex/data-science/
[9] https://marcolorenzi.github.io/teaching.html
[10] http://univ-cotedazur.fr/en/idex/projet-structurant/neuromod

15 laboratories, and aims at integrating interdisciplinary approaches for modeling brain mechanisms and cognitive functions.

- I am currently reviewer of the granting agencies ANR (Agence Nationale de la Recherche, France) and EPSRC (Engineering and Physical Sciences Research Council, UK).

- I am editorial board member of the Neurology Panel of the journal Scientific Reports (Nature Publishing Group) since February 2014, and member of the board of Statisticians of the Journal of Alzheimer's Disease (IOS Press) (2017).

- I am currently reviewer of the granting agencies ANR (Agence Nationale de la Recherche, France) and EPSRC (Engineering and Physical Sciences Research Council, UK).

- I am review of several international conferences, among which: Neural Information Processing Systems (NIPS, since 2016),International Conference on Machine Learning (ICML, since 2017), International Conference on Learning Representations (ICLR, since 2018), Medical Image Computing and Computer Aided Intervention (MICCAI, since 2012), IEEE International Symposium of Biomedical Imaging (ISBI, since 2015), Information Processing in Medical Imaging (IPMI, since 2013), Geometric Science of Information (GSI, 2013-14).

- I am reviewer of scientific journals, among which: Neurobiology of Aging, Alzheimer's Disease & Associated Disorders, Alzheimer's and Dementia, Journal of Alzheimer's Disease, Technology in Cancer Research and Treatment, Medical Image Analysis, IEEE Transactions on Medical Imaging, NeuroImage, International Journal of Computer Vision, Journal of Mathematical Image and Vision, IEEE Transactions on Biomedical Engineering, Human Brain Mapping, Scientific Reports, Neural Computation.

- I have been jury member for the PhD probation exam of several students of Université Côte d'Azur and Eurecom.

- I was invited to present a series of lectures on my work at the summer school *Mathematical Models in Biomedical Imaging*, organized by the University of Granada, Spain, on July 2017.

- A selection of invited lectures include: Martinos Center, MGH, Boston, US (2014), Department of Computing of Imperial College London, London, UK (2014); Brain Imaging Centre, Montreal Neurological Institute, CA (2014); Institut du Cerveau et de la Moelle Épinière, ICM, FR (2016); International Workshop on Geometry, PDE's and Lie Groups in Image Analysis. Eindhoven University of technology, Eindhoven, NL, 2016; Alzheimer's disease strategic roadmap meeting, Geneva, CH, 2019; Collége de France, Paris, FR, 2019; Genoa's University, Genoa, IT, 2019.

Chapter 1

Axis I: Modeling and quantifying temporal changes in biomedical data

This axis concerns the investigation of novel approaches for the probabilistic modeling of spatio-temporal variations in medical images and biological data. My research leverages on the extension of advanced statistical learning methods, such as Gaussian process (GP) regression, for the spatio-temporal analysis of complex and high-dimensional data. The methodological impact of this research project is linked to a series of recent achievements: the prestigious Erbsmann prize awarded to Dr. Sara Garbarino for the work presented to the conference *Information Processing in Medical Imaging* (IPMI 2019) [GLI+19], the oral podium presentation at *International Conference of Machine Learning* (ICML 2018) [LF18], the publication of the monotonic GP regression framework [LFF+17], and its extension to high-dimensional multi-modal imaging data [ANARL19] (both works published in NeuroImage).

1.1 Constrained non-parametric modeling for the analysis of biomedical data

Modern machine learning methods have demonstrated state-of-art performance in representing complex functions in a variety of applications. Nevertheless, the translation of complex learning methods in natural science and in the clinical domain is still challenged by the need of interpretable solutions. To this end, several approaches have been proposed in order to constrain models to plausible forms, such as boundedness [DVM12], monotonicity [RV10], or mechanistic behaviors [ALL13]. This is a crucial requirement to provide a more interpretable and realistic description of natural phenomena [FHP03, ZRB+17].

To tackle the problem of lack of interpretability of data-driven approaches, monotonic regression has recently been proposed to model clinical and biological data [RV10]. Thanks to the monotonic constraint it is indeed possible to identify progression models consistent with the

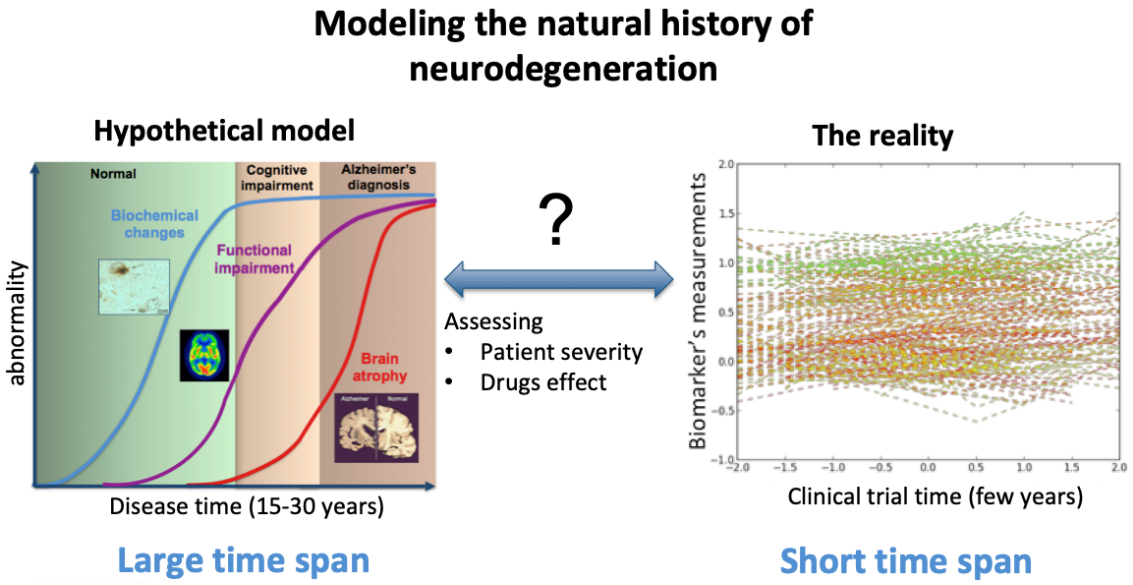


Figure 1.1: Monotonic GP regression allows to estimate models of long term NDD progressions (left) based on the analysis of collections short-term longitudinal measurements available from clinical dataset (right) [LFF+17, GLI+19, NARLar].

evolution of the biomedical measures towards pathological values. I have contributed to this research domain by investigating novel approaches for modeling NDD data through monotonic GP regression [LFF+17], and I showed that by introducing monotonicity in standard GP models it is possible to estimate plausible evolution of the biomarkers over the natural history of the disease (Figure 1.1).

The model is based on the probabilistic estimation of biomarkers' trajectories and on the quantification of the uncertainty of the predicted individual pathological stage. To this end, the inference framework accounts for a time reparameterization function, encoding individual differences in disease timing and speed relative to the fixed effect. Thanks to the probabilistic nature of the GP framework, the resulting long-term disease progression model can be used as a statistical reference representing the transition from normal to pathological stages, thus allowing probabilistic diagnosis in the clinical scenario. The success of this modeling approach led us to the development of a web platform (gpprogressionmodel.inria.fr) to allow clinicians to easily and freely use the software for the analysis of their data. This initiative is currently filed for patent through the *Action de Développement Technologiques (ADT)* at Inria Sophia Antipolis, and used by a number of research partners.

This modeling rationale can be extended to provide comprehensive neurodegeneration models accounting for complex dynamics across measurements. This is a crucial requirement to

provide more realistic description of natural phenomena. For example, bio-physical and mechanistic models are necessary when interpreting and simulating experimental data in biomedical engineering [FHP03, VG07, KRC+11]. However, accounting for the complex properties of biological systems in data-driven modeling approaches poses important challenges. For example, functions are often non-smooth and characterized by nonstationaries which are difficult to encode in shallow models. Complex cases can arise already in classical Ordinary Differential Equations (ODE) systems for certain configurations of the parameters, where functions can exhibit sudden temporal changes [GMM71, Fit55]. In [LFF+17] I generalized the problem of constrained regression to account for a wider set of constraints. These constraints can be expressed through functional relationships on the model dynamics, for example trough ordinary differential equations (ODEs) applied to the derivatives of the GP regression functions.

To tackle this problem, jointly with Prof. Maurizio Filippone (Eurecom, France) I focused on general and flexible GP regression models allowing a rich set of constraints on functions and derivatives of any order. We focused on: i) *inequality constraints*, arising in problems where the class of suitable functions is characterized by specific properties, such as monotonicity or convexity/concavity [RV10]; and ii) *equality constraints* on the function and its derivatives, required when the model should satisfy given physical laws implemented through mechanistic description of a system of interest. A typical example is represented by modeling and inference of ODE systems [MH15].

This latter contribution is at the core of the recent publication with Dr. Sara Garbarino [GLI+19], in which we developed a framework for modeling and inference of protein propagation dynamics from the analysis of brain amyloid imaging data (Figure 1.2). Coherently with the developed theory on constrained regression, protein progression is modelled as a GP, while bio–mechanical processes are imposed as constraints on the proteins dynamics through dynamical ODE systems. The Bayesian setting allows for uncertainty quantification of the proteins dynamics while, to achieve tractability, the inference problem is solved via stochastic variational inference. This framework provides principled theory for model comparison via assessment of model evidence, for example by investigating the Evidence Lower Bound (ELBO) as a surrogate evidence measure. The application to brain amyloid imaging data showed that this approach allows the bio-mechanical interpretation of amyloid deposition in Alzheimer's disease, and achieves accurate predictions of individual amyloid status in unseen data.

1.2 Progression modeling in high-dimensional imaging data

Modeling and prediction of spatio-temporal changes in medical images is limited by important computational challenges. For example the application of non-parameteric learning methods, such as GPs, to the voxel-wise modelling of image time series requires the specification of voxel-by-voxel joint covariance structures, which is in general computationally prohibitive. With the collaboration of Dr. Gabriel Ziegler, I introduced efficient formulations of spatio-temporal GPs

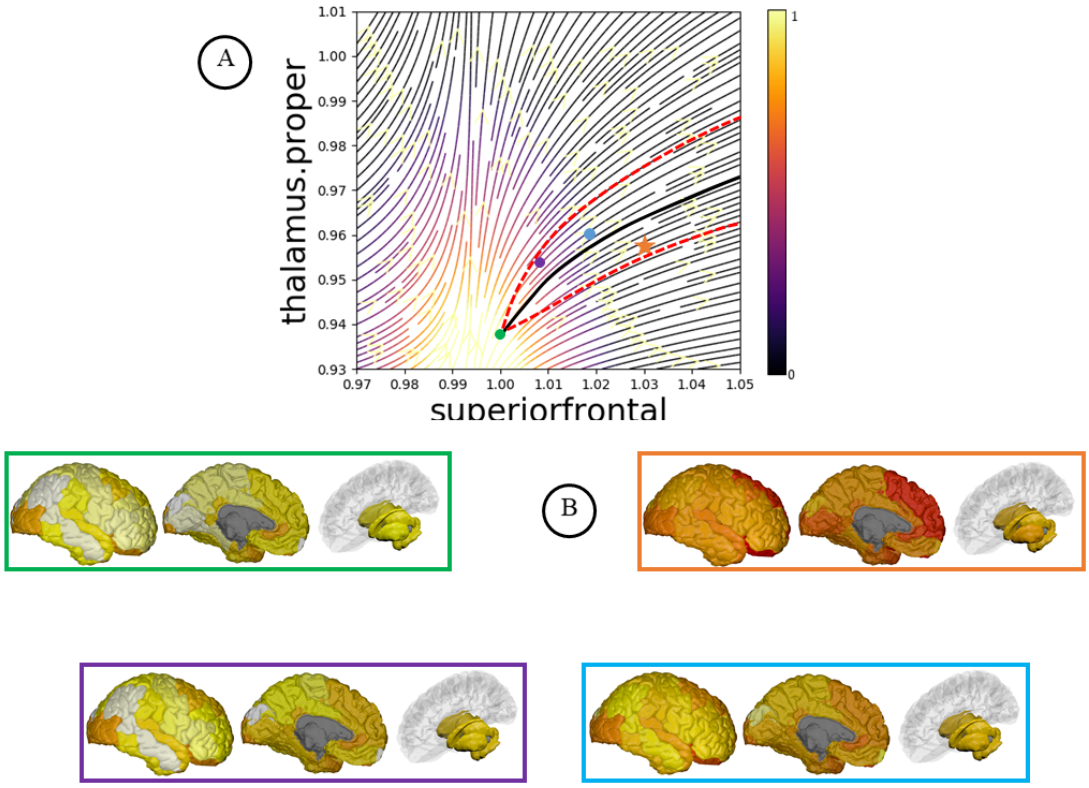


Figure 1.2: Figure from [GLI+19]. A) Streamlines for the estimated amyloid deposition dynamics of superior frontal and thalamus regions for a sample healthy individual converted to MCI at 6 years, and to AD at 8 years. Black lines: predicted dynamics. Colored dots: 3 observed time–points at 4,6,8 years. Star shaped point: unseen follow–up at 10 years; Red dashed lines: associated variability. B) Predicted cumulative amyloid deposition for the time-points highlighted in A).

to model signal changes observed in time-series of medical images [LZAO15]. By proposing a generative model of spatio-temporal signal measured in 3D grids, I provided a flexible and computationally efficient approach for the analysis of aligned image time series by accounting for spatial and temporal correlation. In particular, by assuming local spatial correlation model and separability between spatial and temporal changes, we introduced a very efficient formulation based on a covariance structure parameterized by the Kronecker product of small size covariance matrices. The proposed model extends classical approaches based on parametric models by providing a flexible and efficient statistical tool for the analysis of image features from spatially aligned time series, for instance by allowing statistical inference on the covariance parameters.

This methodology is at the core of the recent work with Clément Abi Nader [NARLar], in which the constrained GP regression is extended to image data, via an efficient generative formulation of disease progression modeling. The key to scalability is based on reformulating the disease model via spatio-temporal matrix factorization, where inference on the sources is constrained by anatomically plausible statistical priors. To model realistic trajectories, the temporal sources are defined as monotonic and time-reparametrized Gaussian Processes. To account for the non-stationarity of brain images, we model the spatial sources as sparse codes convolved at multiple scales. The method has been tested on synthetic data favorably comparing with standard blind source separation approaches. The application on large-scale imaging data from a clinical study allowed us to disentangle differential temporal progression patterns mapping brain regions key to neurodegeneration, while revealing a disease-specific time scale associated to the clinical diagnosis.

This work presents one of the first realizations of a high-dimensional model of Alzheimer's disease progression, jointly encompassing several imaging modalities at full resolution (MRI, amyloid- and fluorodeoxyglucose-PET), along with clinical scores.

Chapter 2

Axis II: Handling heterogeneity in biomedical data

This axis focuses on the definition of learning algorithms for modeling the joint variation between heterogeneous information, such as imaging, clinical, and biological data. The importance and impact of such applications is paramount. The application of statistical learning methods to the growing amount of data available to researchers and clinicians comes with the promise of novel understanding of NDD, along with more effective quantitative measures of the pathology. However, data complexity and heterogeneity represents a major computational and statistical challenge, ultimately affecting interpretability and generalization of the findings.

In my work I defined complementary research strategies to tackle this problem. Firstly, by defining novel validation strategies for multivariate models applied to high-dimensional imaging-genetics data. Secondly, by defining innovative strategies for probabilistic modeling of heterogeneous data sources. These contributions are timely and of high potential. This is demonstrated by the related publications in prestigious scientific journals [LAG+18], as well as by the recent presentation to the conference ICML [AARL19].

2.1 Multivariate latent variable models in imaging-genetics

In genome-wide association studies (GWAS) 1,000,000s of single nucleotide polymorphisms (SNPs) are tested individually for their association with either case-control status [LIVH+13] or disease-specific quantitative phenotypes such as regional brain volumes [PGL+09] or brain amyloid burden [RRN+14]. Due to its ease-of-use and well-established theoretical framework, univariate analysis of genetic data is still the predominant analysis method. However, there are certain statistical and methodological shortcomings such as the massive requirement for multiple testing, redundancies introduced by linkage disequilibrium (LD) and the lack of analysis of epistatic effects (e.g., SNP-SNP interactions), which have to be explicitly modelled and

searched exhaustively [KTAC+12]. In recent years many domains have seen an increased use of multivariate approaches including neuroscience [SRR+13] and GWAS [SHL+10]. Also recent methodological advances in the imaging genetics domain rely on multivariate approaches to capture meaningful genotype-phenotype interactions [SBC+09]. The appeal of these methods lies in their ability to identify complex relationships between the genome and the brain by simultaneously modeling the joint effect of genetic variants on brain features. The promising potential of multivariate imaging-genetics approaches is to explicitly highlight the underlying biology of macroscopic processes, such as brain atrophy, by identifying sets of genetic variants that are jointly associated with phenotype. Multivariate GWAS offer the potential to shed light on the complex genotype-phenotype relationship, and may thus highlight novel links between brain physiology and biological functions. A typical drawback of multivariate imaging-genetics models is related to overfitting: the number of parameters of a multivariate model is usually orders of magnitude higher than the available study sample size, thus biasing the model with respect to the training cohort. Ultimately, the reliability of statistical studies in imaging-genetics, either with GWAS or multivariate approaches, is critically linked to the generalization of the findings to unseen data. The problem is essentially related to the understanding of the functional role of the set of genetic variants, and to the difficulty of replicating statistical results in unseen cohorts.

During my research, with Dr. Andre Altmann (UCL) we focused on the development of statistical analysis approaches based on dimensionality reduction for combining imaging-genetics data. These methods are essentially based on the identification of latent modes of maximal association between different sets of features. The underlying principle consists in looking for pairs of feature combinations - one "combination" or mode for each of the two distinct types - that have maximal statistical association. To this end, I investigated efficient approaches to multivariate analysis based on Partial Least Squares (PLS) for the joint analysis of imaging and genetics data [LAG+18], as well as of different imaging modalities [LSM+16]. In the challenging imaging-genetics case, we tackled the central problem of interpretation and validation of the statistical findings. This issues has an important impact on the understanding of the functional role of sets of genetic variants, and to the difficulty of replicating the statistical results in unseen cohorts. We addressed this technical bottleneck by introducing the idea of *functional prioritization*: high-dimensional modeling results are validated in downstream experiments, through query of high-quality databases of matched genotype and gene expression measurements, such as the Genotype Tissue Expression project GTEx[1] [CAB+15] and BRAINEAC[2] [TRW+11]. In this way we can isolate a low number of functionally relevant SNPs, thus prioritizing few candidate genetic variants hypotheses with a clear functional mechanism, and alleviating the multiple testing burden during the statistical validation.

[1] gtexportal.org/home/
[2] www.braineac.org/

2.2 Modeling multi-channel and multi-organ data

With Luigi Antelmi we studied novel approaches to the joint analysis of heterogeneous data sources represented by multiple information channels (e.g. imaging, clinical and genetics) [AARL19] (Figure 2.1). Our analysis method is based on the extension of modern generative approaches, such as the Variational Autoencoder (VAE) [DW+14].To account for several encoding from multiple-channel information, we imposed a constraint on the latent representations in an information theoretical sense, where each latent representation is enforced to match a common target prior. We showed that this constraint can be optimized within a variational optimization framework, allowing efficient inference of channel encodings and latent representation. Moreover, we showed that the variational framework of VAE can be extended by improving parsimony and interpretability when jointly account for latent relationships across multiple data channels. In the latent space, this is achieved by enforcing parsimonious representations through variational dropout [MAV17]. In extensive synthetic experiments we showed that our model favorably compares to standard non-sparse VAE formulations in presence of multiple data sources. The application to real data showed that the learned dropout parameter can be used for model selection, to automatically identify meaningful latent effect of age and pathology, allowing to predict clinical diagnosis in Alzheimer's Disease. Importantly, due to the general formulation, the proposed method can find various applications as a general data interpretation technique, not limited to the biomedical research area.

Finally, with Jaume Banus Cobo and Dr. Maxime Sermesant we tackled the problem of heterogeneous data integration in case of multi-organ modeling [BLCS19]. In particular, we developed a joint model of brain and heart interaction under biophysical assumptions about the cardiac physiology. This was done by adapting the personalisation of the cardiac parameters of a lumped cardiovascular model under constraints provided by brain features extracted from imaging data (such as ventricles or white matter hyper-intensities volumes). This approach allows to tackle the ill-posed nature of the personalisation, and to identify plausible solutions across the population. The optimized of such a model relies on a genetic optimization algorithm, equipped with a regularisation term taking into account brain features as additional constraint. We applied this framework to a large cohort composed by more than 3000 subjects for which cardiac and brain information was jointly available in the UK Biobank. This approach allowed us to identify statistically significant associations between the personalised model parameters and brain volumetric features in specific cardiovascular conditions, such as atrial fibrillation, that match findings reported in previous clinical studies. In the extension of this work we aim at modeling the local blood flow in the brain, and studying the relationship between estimated brain vascular parameters and NDD.

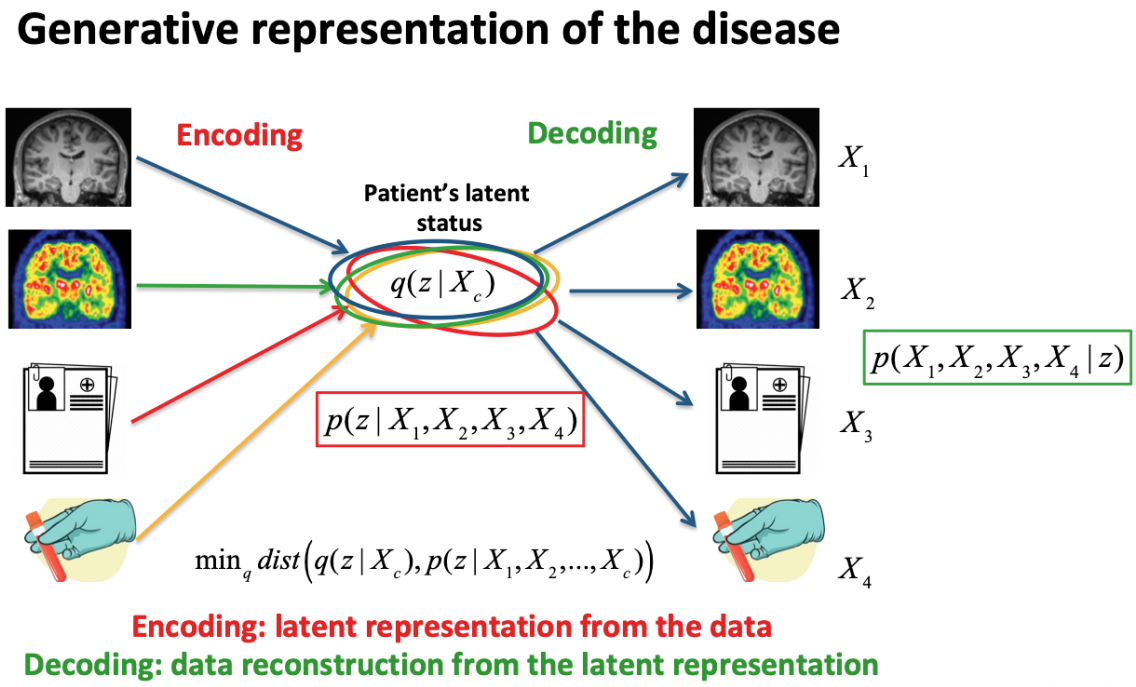


Figure 2.1: Axis II. Multi-channel variational autoencoder (McVAE). To enforce a consistent representation of multimodal information, latent encoding from different channels are constrained to match a common target posterior distribution.

Chapter 3

Axis III: Federated statistical learning for meta-analyses of biomedical data

This research axis represents a novel direction of investigation in which I envision the extension of the methodology proposed in the previous chapters to cope with the constraint of data privacy and security in multi-centric studies. In this chapter I introduce the rationale behind the recently funded projects on this topic, that will start during the next year.

Databanks worldwide currently contain biomedical information in previously unimaginable numbers. Fostered by recent advances in statistical learning, this is creating great enthusiasm around the explosion of "Big Data" in healthcare. Machine learning-based methods are for the first time at the core of FDA approved systems for automated retinopathy detection, and have shown human-like level diagnostic accuracy in challenging diagnostic tasks, for example for tumour detection in mammography. These results have been obtained thanks to unique organizational and economical efforts for the establishment of large-scale centralized data collection initiatives: the key component of machine learning-based system is the availability of very large training databases.

This is however far from being the standard working scenario in healthcare: the effective exploitation of modern machine learning methods poses tough methodological and translational challenges, since the predominant paradigm consists in highly secured and private datasets usually providing information for low- or middle-sized data samples (from tens to hundreds of cases). The analysis of biomedical data must indeed adhere to very strict specificities: anonymity, security, and non-transferability of data across centers [ADW+18]. This problem has been recently exacerbated by the strict regulations imposed around data protection, such as the EU "General Data Protection Regulation" (GDPR). When imposing these limitations, the majority of current powerful learning methods simply do not find a straightforward application.

In healthcare, the concept of "model federation" has long been applied in the context of meta-analyses of multi-centric studies. The underlying rationale of meta-analysis is that powered estimations of test statistics or effect-sizes can be obtained by pooling together estimates from independent individuals studies, while accounting for the associated uncertainty. Standard approaches to meta-analysis essentially rely on univariate testing, and long-term and consolidated amount of literature is today available on the topic, especially in the field of epidemiology [RGL+08, WHK08, SP04]. Based on this rationale, the ENIGMA[1] consortium led to important imaging-genetics findings on typical sample sizes of 30,000 or more individuals. However, when the features to be analysed are in the order of millions (e.g. in case of genetics variants or brain images), the mass-univariate paradigm is prone to statistical limitations, such as the multiple comparisons problem, as well as interpretability issues when features are highly correlated. Ultimately, much of the implicit underpinnings revealed by this sort of analysis remain elusive: a single variant at best has only a tiny effect on brain variation. All in all, mass-univariate results often lack of stability and reproducibility.

Only recently, multivariate statistical methods have been formulated in a federated context. These advances are grounded on the separability and linearity properties of standard multivariate modelling functionals. These properties allow to "split" a functional across data instances, in such a way that an optimization task can be distributed and subsequently merged across data partitions. This is the case for a wide range of standard methods based on least squares cost and linear transformations. As a result, we can today rely on federated counterparts of multivariate regression models [HSR+15], feature selection [GT14], clustering [CM09], and dimensionality reduction techniques [MG03, GS13].

In neuroimaging, few recent works attempted distributed analysis mostly via linear modelling and classical dimensionality reduction [DBM+13, JSM+12, MVS+17, BSC+15]. For example, a large body of optimization problems of the kind $\hat{\mathbf{W}} = \text{argmin}_{\mathbf{W}} \mathcal{L}(\mathbf{X}, Y|\mathbf{W})$ can be federated as a constrained optimization task:

$$\hat{\mathbf{W}} = argmin_{\mathbf{W}} \sum_i argmin_{\mathbf{V}_i} \mathcal{L}(\mathbf{X}_i, Y_i|\mathbf{V}_i) + \mathcal{H}(\mathbf{V}_i, \mathbf{W}),$$

where the parameters $\mathbf{V}_i$ are local solutions to the problem $\mathcal{L}$, while the term $\mathcal{H}$ is a constraint enforcing each local solution towards a common term $\mathbf{W}$. The solution $\hat{\mathbf{W}}$ thus represents the federated counterpart of each local optimization problem. Under convexity and separability constraints on the functionals $\mathcal{L}$ and $\mathcal{H}$, this kind of problem can be efficiently optimized, for example via the augmented Lagrangian formulation of the Alternating Direction Method of Multipiers (ADMM) [BPC+11]. Our recent initiative Meta-ImaGen is in continuity with these efforts [SGR+19], and provides a pipeline for data standardization, linear modelling and dimensionality reduction.

[1] http://enigma.ini.usc.edu

The state of the art in federated learning for complex models such as DNN is based on stochastic gradient descent and mini-batch optimization, and leverages on the linearity properties of the cost gradient. When a cost functional is separable across data instances, such as with a squared loss function, the gradients of the parameters can be collected across different centers and combined to form an update for a centralized model. Based on this setup, a number of approaches have been proposed, such as in [MMR+16, JBHS17]. The straightforward implementation of gradient-based strategies for federated learning approaches presents however important computational and methodological bottlenecks.

First, the number of iterations required by gradient-based optimization, in the order of thousands for a standard neural network, does not allow to easily deploy such a strategy when gradients have to be transmitted between remote nodes. Although recent extensions aim at mitigating this problem by limiting the number of exchanges across centers, this approach still presents important computational challenges when applied to large-dimensional data, such as the biomedical one.

Second, model's prediction and parameters may sensitively differ when trained on different data instances.This is a very common weakness in machine learning, related to model variability and data bias. The over-representation of specific traits in a data instance (i.e. non iid distribution of the data across instances) may likely influence the outcome of a model, and thus introduce a bias reflecting the specificity of the training data. These issues may have important adverse effects in federated learning, as they may expose the system to failures in preserving data anonymity, as well as to lack in robustness.

To tackle these problems, I aim at developing a principled theoretical framework to define and automatically identify bias in federated learning systems. These challenges can be effectively tackled by leveraging on variational inference in a Bayesian non-parametric setting.

3.1 Towards federated Bayesian non-parametric analysis of multi-centric studies

To the best of our knowledge, only recently researchers have been turning their attention to the federated formulation of Bayesian methods in biomedical data analysis [ABBB19, MGDAK18]. Adapting GP, and more generally Bayesian non-parametric methods, to federated analysis is a critical requirement for extending the discovery potential in today's large-scale data scenario.

The starting point of this research avenue is our consolidated research line on statistical learning, and biomedical data analysis. Our currently available software for federated analysis already provides a basic federated analysis toolbox based on the extension of standard algorithms. We aim at enriching the toolbox with a wide range of federated probabilistic models, based on the extension of our current works on Bayesian non-parametric methods. In particular, in the ANR funded project Fed-BioMed we aim at the application of our methodology on clinical population of unprecedented dimensions for this kind of study, including several databases

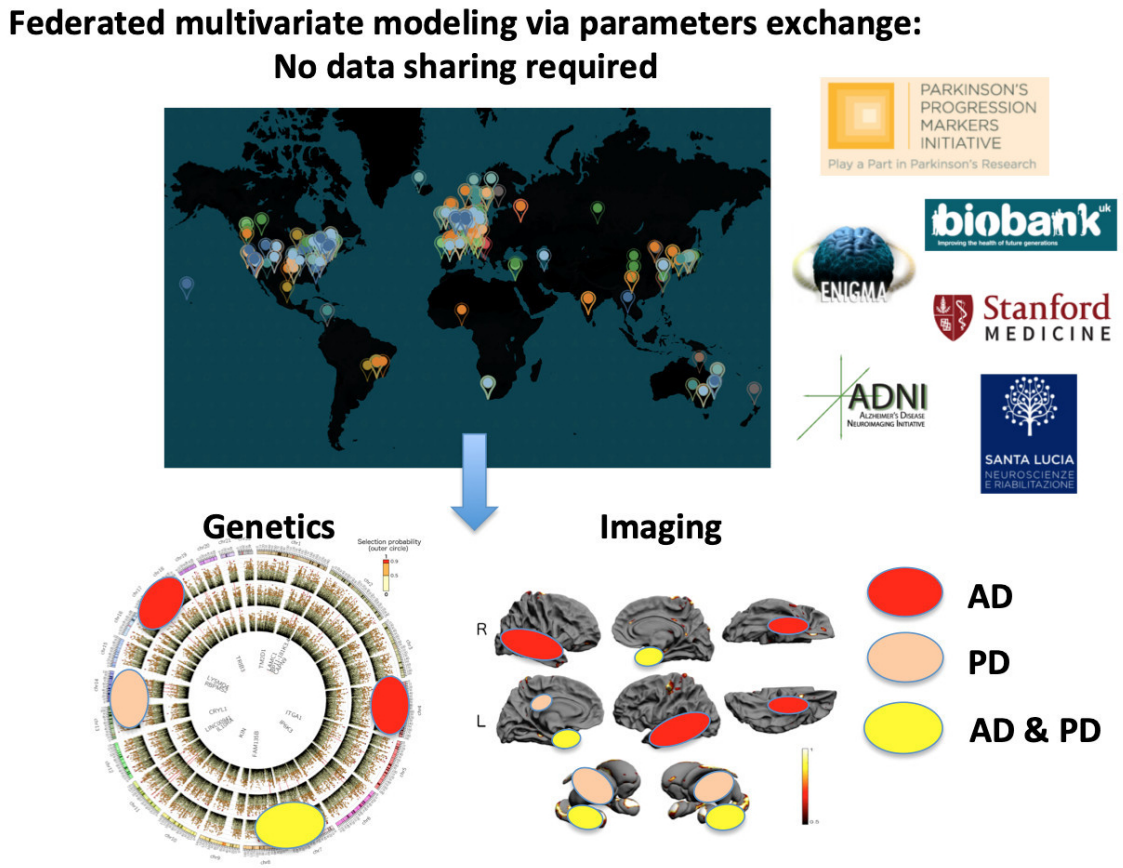


Figure 3.1: Axis III. Multi-centric imaging-genetics application in the upcoming project Fed-BioMed. AD: Alzheimer's disease; PD: Parkinson's disease

representing the full spectrum of variability across healthy condition, Alzheimer's disease (AD) and Parkinson's Disease (PD). Our project will access the data through research partnerships (ENIGMA consortium, Stanford University, US, and Fondazione Santa Lucia, IT), or thanks to formal applications (ADNI, PPMI, UK Biobank) (Figure 3.1).

Finally, by reformulating stochastic latent variable models in a federated setting, with Dr. Maxime Sermesant we will analyze the variability of cardiac imaging data for 1200+ cases across cohorts and disease status from a multi-centric study planned by the Bordeaux University Hospital. The resulting federated model will allow better understanding and probabilistic prediction of sudden cardiac death in the clinical practice.

This line of research has the potential to provide multiple contributions to the state-of-the-art. From the methodological perspective, it will require to study novel approaches to

federate distribution parameters within our Bayesian setting, accounting for bias and unbalanced datasets. Current inference schemes, such as based on variational methods, must be reformulated to accommodate for the distributed and non-iid nature of the data. We will also need to investigate novel optimization schemes across networks, along with the associated computational cost.

Chapter 4

Conclusions and Perspectives

During the past years I consolidated my experience in the analysis of biomedical data with privileged application to the study of neurodegenerative pathologies. Thanks to the collaborative network developed during my research, I am currently pursuing an investigative program in strict contact with research, clinical, and industry partners, tackling central problems in current applications of data science in healthcare.

Modeling disease progression and temporal changes in biomedical data is an open challenge, and a central aim of my research project is to advance the state-of-art in this field. In spite of recent methodological advances in the analysis of NDD data, much still has to be done in terms of validation and generalization of current approaches for their use in the clinical scenario. Current research cohorts are still subject to several forms of bias, depending for example on specific enrollment and data acquisition strategies. In this case, models themselves are subjected to this kind of bias, and prone to overfit. For this reason, more research is needed to increase robustness and generalization of current approaches. For the same reason, model scalability plays a central role and should represent a primary development aspect. The increasing complexity of the targeted biological processes, data dimensionality and sample size will require to researchers a continuously increasing scientific effort. The availability of biomedical measures, along with the large amount of data that can be acquired with novel technologies such as wearable devices and serious games [KSB+15], will require novel and flexible methods to account for the complexity and high-resolution of this information.

We are witnessing major technical advancements related to the introduction of modern machine learning tools for computing and optimization, such as PyTorch [PGC+17], or TensorFlow [AAB+15]. Automatic gradient computation based on back-propagation is becoming a strategic technique for increasing the flexibility and ease of development of current models. More research can be done in this sense, to expand the potential of these approaches to introduce even more complex modeling hypothesis, such as for consistently accounting for non-parametric dynamics [CRBD18]. This concept is in line with my research on progression

modeling constrained by dynamical systems, and should be further expanded to account for partial derivatives with complex support, such as in 3D manifolds.

More generally, the study of computationally tractable approaches for Bayesian modeling in heterogeneous data finds important application is several domains, beyond health-care and biostatistics. Flexibility and generalization to different modeling problems is among the priorities of my work. This is a key aspect to maximize the impact of research on society, through development of spin-off projects and industry transfer.

Finally, the definition of federated learning strategies for the analysis of biomedical data will introduce a novel practice in the domain, and will allow to scale current modeling capabilities to data of previously unimaginable size. This kind of research has the potential to create a novel standard for biomedical data analysis in the upcoming era of large scale studies.

Publications list

Here I list the ensemble of work published after my PhD.

Journal Papers

[1] **Marco Lorenzi**, Andre Altmann, Boris Gutman, Selina Wray, Charles Arber, Derrek P Hibar, Neda Jahanshad, Jonathan M Schott, Daniel C Alexander, Paul M Thompson, et al. Susceptibility of brain atrophy to trib3 in alzheimer's disease, evidence from functional prioritization in imaging genetics. *Proceedings of the National Academy of Sciences*, 115(12):3162–3167, 2018.

[2] **Marco Lorenzi**, Nicholas Ayache, Xavier Pennec, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Regional flux analysis for discovering and quantifying anatomical changes: an application to the brain morphometry in alzheimer's disease. *Neuroimage*, 115:224–234, 2015.

[3] **Marco Lorenzi**, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, Sébastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease. *NeuroImage*, 2017.

[4] **Marco Lorenzi**, Xavier Pennec, Giovanni B Frisoni, Nicholas Ayache, Alzheimer's Disease Neuroimaging Initiative, et al. Disentangling normal aging from alzheimer's disease in structural magnetic resonance images. *Neurobiology of aging*, 36:S42–S52, 2015.

[5] **Marco Lorenzi**, Ivor J Simpson, Alex F Mendelson, Sjoerd B Vos, M Jorge Cardoso, Marc Modat, Jonathan M Schott, and Sebastien Ourselin. Multimodal image analysis in alzheimer's disease via statistical modelling of non-local intensity correlations. *Scientific reports*, 6:22161, 2016.

[6] David M Cash, Chris Frost, Leonardo O Iheme, Devrim Ünay, Melek Kandemir, Jurgen Fripp, Olivier Salvado, Pierrick Bourgeat, Martin Reuter, Bruce Fischl, et al. Assessing

atrophy measurement techniques in dementia: results from the miriad atrophy challenge. *Neuroimage*, 123:149–164, 2015.

[7] Claire Cury, Stanley Durrleman, David M Cash, **Marco Lorenzi**, Jennifer M Nicholas, Martina Bocchetta, John C Van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, et al. Spatiotemporal analysis for detection of pre-symptomatic shape changes in neurodegenerative diseases: Initial application to the genfi cohort. *NeuroImage*, 188:282–290, 2019.

[8] Sebastiano Ferraris, Johannes Van Der Merwe, Lennart Van Der Veeken, Ferran Prados, Juan-Eugenio Iglesias, Andrew Melbourne, **Marco Lorenzi**, Marc Modat, Willy Gsell, Jan Deprest, et al. A magnetic resonance multi-atlas for the neonatal rabbit brain. *NeuroImage*, 179:187–198, 2018.

[9] Mehdi Hadj-Hamou, **Marco Lorenzi**, Nicholas Ayache, and Xavier Pennec. Longitudinal analysis of image time series with diffeomorphic deformations: a computational framework based on stationary velocity fields. *Frontiers in neuroscience*, 10:236, 2016.

[10] Bishesh Khanal, **Marco Lorenzi**, Nicholas Ayache, and Xavier Pennec. A biophysical model of brain deformation to simulate and analyze longitudinal mris of patients with alzheimer's disease. *NeuroImage*, 134:35–52, 2016.

[11] Alex F Mendelson, Maria A Zuluaga, **Marco Lorenzi**, Brian F Hutton, Sébastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Selection bias in the reported performances of ad classification pipelines. *NeuroImage: Clinical*, 14:400–416, 2017.

[12] Clément Abi Nader, Nicholas Ayache, Philippe Robert, and **Marco Lorenzi**. Monotonic gaussian process for spatio-temporal trajectory separation in brain imaging data. *NeuroImage*, 2019, to appear.

[13] Antonio Preti, Cristina Muscio, Marina Boccardi, **Marco Lorenzi**, Giovanni De Girolamo, and Giovanni Frisoni. Impact of alcohol consumption in healthy adults: a magnetic resonance imaging investigation. *Psychiatry Research: Neuroimaging*, 224(2):96–103, 2014.

[14] Marzia A Scelsi, Raiyan R Khan, **Marco Lorenzi**, Leigh Christopher, Michael D Greicius, Jonathan M Schott, Sebastien Ourselin, and Andre Altmann. Genetic study of multimodal imaging alzheimer's disease progression score implicates novel loci. *Brain*, 141(7):2167–2180, 2018.

[15] Raphaël Sivera, Hervé Delingette, **Marco Lorenzi**, Xavier Pennec, Nicholas Ayache, Alzheimer's Disease Neuroimaging Initiative, et al. A model of brain morphological changes related to aging and alzheimer's disease from cross-sectional assessments. *NeuroImage*, 2019.

Conferences with proceedings

[1] **Marco Lorenzi** and Maurizio Filippone. Constraining the dynamics of deep probabilistic models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3227–3236, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[2] **Marco Lorenzi**, Boris Gutman, Derrek P Hibar, Andre Altmann, Neda Jahanshad, Paul M Thompson, and Sebastien Ourselin. Partial least squares modelling for imaging-genetics in alzheimer's disease: Plausibility and generalization. In *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 838–841. IEEE, 2016.

[3] **Marco Lorenzi**, Boris Gutman, Paul M Thompson, Daniel C Alexander, Sebastien Ourselin, and Andre Altmann. Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study. In *12th International Symposium on Medical Information Processing and Analysis*, volume 10160, page 1016016. International Society for Optics and Photonics, 2017.

[4] **Marco Lorenzi**, Gabriel Ziegler, Daniel C Alexander, and Sebastien Ourselin. Efficient gaussian process-based modelling and prediction of image time series. In *International Conference on Information Processing in Medical Imaging*, pages 626–637. Springer, 2015.

[5] **Marco Lorenzi**, Gabriel Ziegler, Daniel C Alexander, and Sebastien Ourselin. Modelling non-stationary and non-separable spatio-temporal changes in neurodegeneration via gaussian process convolution. In *Medical Learning Meets Medical Imaging*, pages 35–44. Springer, 2015.

[6] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and **Marco Lorenzi**. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 302–311, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[7] Jaume Banus Cobo, **Marco Lorenzi**, Oscar Camara, and Maxime Sermesant. Large scale cardiovascular model personalisation for mechanistic analysis of heart and brain interactions. In *Proceedings of the 10th International Conference on Functional Imaging and Modeling of the Heart (FIMH)*, Bordeaux, France, Jun 2019.

[8] Sebastiano Ferraris, **Marco Lorenzi**, Pankaj Daga, Marc Modat, and Tom Vercauteren. Accurate small deformation exponential approximant to integrate large velocity fields:

Application to image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24, 2016.

[9] Sara Garbarino and **Marco Lorenzi**. Modeling and inference of spatio-temporal protein dynamics across brain networks. In *International Conference on Information Processing in Medical Imaging*, pages 57–69. Springer, 2019.

[10] Boris A Gutman, P Thomas Fletcher, M Jorge Cardoso, Greg M Fleishman, **Marco Lorenzi**, Paul M Thompson, and Sebastien Ourselin. A riemannian framework for intrinsic comparison of closed genus-zero shapes. In *International Conference on Information Processing in Medical Imaging*, pages 205–218. Springer, 2015.

[11] Juan Eugenio Iglesias, **Marco Lorenzi**, Sebastiano Ferraris, Loïc Peter, Marc Modat, Allison Stevens, Bruce Fischl, and Tom Vercauteren. Model-based refinement of nonlinear registrations in 3d histology reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 147–155. Springer, 2018.

[12] Bishesh Khanal, **Marco Lorenzi**, Nicholas Ayache, and Xavier Pennec. A biophysical model of shape changes due to atrophy in the brain with alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 41–48. Springer, 2014.

[13] Giorgos Lazaridis, **Marco Lorenzi**, and David Garway-Heat. Enhancing OCT signal by fusion of gans: Improving statistical power of glaucoma trials. In *International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI*. Springer, 2019.

[14] Răzvan Valentin Marinescu, Arman Eshaghi, **Marco Lorenzi**, Alexandra L. Young, Neil P. Oxtoby, Sara Garbarino, Timothy J. Shakespeare, Sebastian J. Crutch, and Daniel C. Alexander. A vertex clustering model for disease progression: Application to cortical thickness images. In *Information Processing in Medical Imaging*, pages 134–145. Springer International Publishing, 2017.

[15] Eliza Orasanu, Pierre-Louis Bazin, Andrew Melbourne, **Marco Lorenzi**, Herve Lombaert, Nicola J. Robertson, Giles Kendall, Nikolaus Weiskopf, Neil Marlow, and Sebastien Ourselin. Longitudinal analysis of the preterm cortex using multi-modal spectral matching. In *Medical Image Computing and Computer-Assisted Intervention*, pages 255–263. Springer International Publishing, 2016.

[16] Santiago Silva, Boris Gutman, Eduardo Romero, Paul Thompson, Andre Altmann, and **Marco Lorenzi**. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, Venice, Italy, Apr 2019.

References

[AAB+15] Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[AARL19] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 302–311, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[ABBB19] John Ashburner, Mikael Brudfors, Kevin Bronik, and Yaël Balbastre. An algorithm for learning shape and appearance models without annotations. *Medical image analysis*, 55:197, 2019.

[ADW+18] April Moreno Arellano, Wenrui Dai, Shuang Wang, Xiaoqian Jiang, and Lucila Ohno-Machado. Privacy policy and technology in biomedical data science. *Annual review of biomedical data science*, 1:115–129, 2018.

[AF00b] John Ashburner and Karl J Friston. Voxel-based morphometry - the methods. *Neuroimage*, 11(6):805–821, 2000.

[ALL13] Mauricio A Alvarez, David Luengo, and Neil D Lawrence. Linear latent force models using Gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2693–2705, 2013.

[ANARL19] Clément Abi Nader, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Monotonic gaussian process for spatio-temporal disease progression modeling in brain imaging data. *NeuroImage*, 2019.

[BLCS19] Jaume Banus, Marco Lorenzi, Oscar Camara, and Maxime Sermesant. Large scale cardiovascular model personalisation for mechanistic analysis

of heart and brain interactions. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 285–293. Springer, 2019.

[BPC+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[BSC+15] Bradley T Baker, Rogers F Silva, Vince D Calhoun, Anand D Sarwate, and Sergey M Plis. Large scale collaboration with autonomy: Decentralized data ica. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015.

[BWSL04] Ella Bossy-Wetzel, Robert Schwarzenbacher, and Stuart A Lipton. Molecular pathways to neurodegeneration. *Nature medicine*, 10(7s):S2, 2004.

[C+15] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[CAB+15] Latarsha J Carithers, Kristin Ardlie, Mary Barcus, Philip A Branton, Angela Britton, Stephen A Buia, Carolyn C Compton, David S DeLuca, Joanne Peter-Demchok, Ellen T Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and biobanking*, 13(5):311–319, 2015.

[CM09] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in neural information processing systems*, pages 289–296, 2009.

[CRBD18] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[CSGC16] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392, 2016.

[Dav04] Christos Davatzikos. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *Neuroimage*, 23(1):17–20, 2004.

[DBM+13] Dany Doiron, Paul Burton, Yannick Marcon, Amadou Gaye, Bruce HR Wolffenbuttel, Markus Perola, Ronald P Stolk, Luisa Foco, Cosetta Minelli,

Melanie Waldenberger, et al. Data harmonization and federated analysis of population-based studies: the bioshare project. *Emerging themes in epidemiology*, 10(1):12, 2013.

[DFJ+14] Bruno Dubois, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, et al. Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria. *The Lancet Neurology*, 13(6):614–629, 2014.

[Dud13] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

[DVM12] Sébastien Da Veiga and Amandine Marrel. Gaussian process modeling with inequality constraints. *Annales de la faculté des sciences de Toulouse Mathématiques*, 21(3):529–555, April 2012.

[DW+14] P Kingma Diederik, Max Welling, et al. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[FHP03] Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[Fit55] Richard FitzHugh. Mathematical models of threshold phenomena in the nerve membrane. *The bulletin of mathematical biophysics*, 17(4):257–278, 1955.

[GLI+19] Sara Garbarino, Marco Lorenzi, AlzheimerâĂŹs Disease Neuroimaging Initiative, et al. Modeling and inference of spatio-temporal protein dynamics across brain networks. In *International Conference on Information Processing in Medical Imaging*, pages 57–69. Springer, 2019.

[GMM71] Narendra S Goel, Samaresh C Maitra, and Elliott W Montroll. On the volterra and other nonlinear models of interacting populations. *Reviews of modern physics*, 43(2):231, 1971.

[GS13] Zhiqiang Ge and Zhihuan Song. Distributed pca model for plant-wide process monitoring. *Industrial & Engineering Chemistry Research*, 52(5):1947–1957, 2013.

[GT14] Samuel M Gross and Robert Tibshirani. Collaborative regression. *Biostatistics*, 16(2):326–338, 2014.

[HMB+16] Marie-Odile Habert, Sullivan Marie, Hugo Bertin, Moana Reynal, Jean-Baptiste Martini, Mamadou Diallo, Aurélie Kas, Régine Trébossen, et al. Optimization of brain pet imaging for a multicentre trial: the french cati experience. *EJNMMI physics*, 3(1):6, 2016.

[HSR+15] Derrek P Hibar, Jason L Stein, Miguel E Renteria, Alejandro Arias-Vasquez, Sylvane Desrivières, Neda Jahanshad, Roberto Toro, Katharina Wittfeld, Lucija Abramovic, Micael Andersson, et al. Common genetic variants influence human subcortical brain structures. *Nature*, 520(7546):224, 2015.

[JBHS17] Zhanhong Jiang, Aditya Balu, Chinmay Hegde, and Soumik Sarkar. Collaborative deep learning in fixed topology networks. In *Advances in Neural Information Processing Systems*, pages 5904–5914, 2017.

[JSM+12] EM Jones, NA Sheehan, N Masca, SE Wallace, MJ Murtagh, and PR Burton. Datashield–shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk epidemiologi*, 21(2), 2012.

[KRC+11] Ender Konukoglu, Jatin Relan, Ulas Cilingir, Bjoern H Menze, Phani Chinchapatnam, Amir Jadidi, Hubert Cochet, Meleze Hocini, Hervé Delingette, Pierre Jaïs, et al. Efficient probabilistic model personalization integrating uncertainty on data and parameters: Application to eikonal-diffusion models in cardiac electrophysiology. *Progress in biophysics and molecular biology*, 107(1):134–146, 2011.

[KSB+15] Alexandra König, Guillaume Sacco, Gregory Bensadoun, Francois Bremond, Renaud David, Frans Verhey, Pauline Aalten, Philippe Robert, and Valeria Manera. The role of information and communication technologies in clinical trials with patients with alzheimerâĂŹs disease and related disorders. *Frontiers in aging neuroscience*, 7:110, 2015.

[KTAC+12] Tony Kam-Thong, Chloé-Agathe Azencott, Lawrence Cayton, Benno Pütz, André Altmann, Nazanin Karbalai, Philipp G Sämann, Bernhard Schölkopf, Bertram Müller-Myhsok, and Karsten M Borgwardt. Glide: Gpu-based linear regression for detection of epistasis. *Human heredity*, 73(4):220–236, 2012.

[LAG+18] Marco Lorenzi, Andre Altmann, Boris Gutman, Selina Wray, Charles Arber, Derrek P Hibar, Neda Jahanshad, Jonathan M Schott, Daniel C Alexander, Paul M Thompson, et al. Susceptibility of brain atrophy to trib3 in alzheimerâĂŹs disease, evidence from functional prioritization in imaging genetics. *Proceedings of the National Academy of Sciences*, 115(12):3162–3167, 2018.

[LF18] Marco Lorenzi and Maurizio Filippone. Constraining the dynamics of deep probabilistic models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3227–3236, StockholmsmÃďssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[LFF+17] Marco Lorenzi, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, Sébastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease. *NeuroImage*, 2017.

[LIVH+13] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452, 2013.

[LSM+16] Marco Lorenzi, Ivor J Simpson, Alex F Mendelson, Sjoerd B Vos, M Jorge Cardoso, Marc Modat, Jonathan M Schott, and Sebastien Ourselin. Multimodal image analysis in alzheimerâĂŹs disease via statistical modelling of non-local intensity correlations. *Scientific reports*, 6:22161, 2016.

[LZAO15] Marco Lorenzi, Gabriel Ziegler, Daniel C Alexander, and Sebastien Ourselin. Efficient gaussian process-based modelling and prediction of image time series. In *International Conference on Information Processing in Medical Imaging*, pages 626–637. Springer, 2015.

[MAV17] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2498–2507. JMLR. org, 2017.

[MG03] Srujana Merugu and Joydeep Ghosh. Privacy-preserving distributed clustering using generative models. In *Third IEEE International Conference on Data Mining*, pages 211–218. IEEE, 2003.

[MGDAK18] Lester Melie-Garcia, Bogdan Draganski, John Ashburner, and Ferath Kherif. Multiple linear regression: Bayesian inference for distributed and big data in the medical informatics platform of the human brain project. jan 2018.

[MH15] Benn Macdonald and Dirk Husmeier. Gradient Matching Methods for Computational Inference in Mechanistic Models for Systems Biology: A Review

and Comparative Analysis. *Frontiers in Bioengineering and Biotechnology*, 3:180, 2015.

[MMR+16] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2016.

[MVS+17] Jing Ming, Eric Verner, Anand Sarwate, Ross Kelly, Cory Reed, Torran Kahleck, Rogers Silva, Sandeep Panta, Jessica Turner, Sergey Plis, et al. Coinstac: Decentralizing the future of brain imaging analysis. *F1000Research*, 6, 2017.

[Nac01] Michael W Nachman. Single nucleotide polymorphisms and recombination rate in humans. *TRENDS in Genetics*, 17(9):481–485, 2001.

[NARLar] Clément Abi Nader, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Monotonic gaussian process for spatio-temporal trajectory separation in brain imaging data. *NeuroImage*, 2019, to appear.

[PAB+10] Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010.

[PGC+17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[PGL+09] Steven G Potkin, Guia Guffanti, Anita Lakatos, Jessica A Turner, Frithjof Kruggel, James H Fallon, Andrew J Saykin, Alessandro Orro, Sara Lupoli, Erika Salvi, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for alzheimer's disease. *PloS one*, 4(8):e6501, 2009.

[RGL+08] Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

[RRN+14] Vijay K Ramanan, Shannon L Risacher, Kwangsik Nho, Sungeun Kim, Shanker Swaminathan, Li Shen, Tatiana M Foroud, Hakon Hakonarson, Matthew J Huentelman, Paul S Aisen, et al. Apoe and bche as modulators of cerebral amyloid deposition: a florbetapir pet genome-wide association study. *Molecular psychiatry*, 19(3):351, 2014.

[RV10] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In Yee W. Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 645–652, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR.

[SACD15] Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems*, pages 2404–2412, 2015.

[SBC+09] Silke Szymczak, Joanna M Biernacka, Heather J Cordell, Oscar González-Recio, Inke R König, Heping Zhang, and Yan V Sun. Machine learning in genome-wide association studies. *Genetic epidemiology*, 33(S1):S51–S57, 2009.

[SGA+15] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[SGR+19] Santiago Silva, Boris Gutman, Eduardo Romero, Paul Thompson, Andre Altmann, and Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, Venice, Italy, Apr 2019.

[SHL+10] Jason L Stein, Xue Hua, Suh Lee, April J Ho, Alex D Leow, Arthur W Toga, Andrew J Saykin, Li Shen, Tatiana Foroud, Nathan Pankratz, et al. Voxelwise genome-wide association study (vGWAS). *Neuroimage*, 53(3):1160–1174, 2010.

[SJH+12] Matt Silver, Eva Janousova, Xue Hua, Paul M Thompson, Giovanni Montana, Alzheimer's Disease Neuroimaging Initiative, et al. Identification of gene pathways implicated in alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, 63(3):1681–1694, 2012.

[SN18] Stephen M Smith and Thomas E Nichols. Statistical challenges in âĂIJbig dataâĂİ human neuroimaging. *Neuron*, 97(2):263–268, 2018.

[SP04] RJS Simpson and Karl Pearson. Report on certain enteric fever inoculation statistics. *The British Medical Journal*, pages 1243–1246, 1904.

[SRR+13] Jessica Schrouff, Maria J Rosa, Jane M Rondina, Andre F Marquand, Carlton Chu, John Ashburner, Christophe Phillips, Jonas Richiardi, and Janaina Mourão-Miranda. PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, 11(3):319–337, 2013.

[The17] Lancet The. Dementia burden coming into focus. *Lancet (London, England)*, 390(10113):2606, 2017.

[TRW+11] Daniah Trabzuni, Mina Ryten, Robert Walker, Colin Smith, Sabaena Imran, Adaikalavan Ramasamy, Michael E Weale, and John Hardy. Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *Journal of neurochemistry*, 119(2):275–282, 2011.

[TWT18] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581, 2018.

[VG07] Vladislav Vyshemirsky and Mark A Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2007.

[WHK08] Esteban Walker, Adrian V Hernandez, and Michael W Kattan. Meta-analysis: Its strengths and limitations. *Cleveland Clinic journal of medicine*, 75(6):431, 2008.

[YFL+11] Eric Yang, Michael Farnum, Victor Lobanov, Tim Schultz, Nandini Raghavan, Mahesh N Samtani, Gerald Novak, Vaibhav Narayan, Allitia DiBernardo, Alzheimer's Disease Neuroimaging Initiative, et al. Quantifying the pathophysiological timeline of alzheimer's disease. *Journal of Alzheimer's Disease*, 26(4):745–753, 2011.

[YHG+17] Jiarui Yang, Chenhui Hu, Ning Guo, Joyita Dutta, Lucia M Vaina, Keith A Johnson, Jorge Sepulcre, Georges El Fakhri, and Quanzheng Li. Partial volume correction for pet quantification and its impact on brain network in alzheimerâĂŹs disease. *Scientific reports*, 7(1):13035, 2017.

[ZRB+17] Gabriel Ziegler, Gerard R Ridgway, Sarah-Jayne Blakemore, John Ashburner, and Will Penny. Multivariate dynamical modelling of structural change during development. *NeuroImage*, 147:746–762, 2017.

Appendices

Selected publications for Axis I

Modeling and inference of spatio-temporal protein dynamics across brain networks

Sara Garbarino[1] and Marco Lorenzi[1]
1. EPIONE project-team, INRIA, Université Côte d'Azur, Sophia Antipolis, France

Originally published in:
Information Processing in Medical Imaging (IPMI), 57-69, 2019

Abstract

Many neurodegenerative disorders (NDs), including Alzheimer's disease, Parkinson's disease, and Huntington's disease, are proteinopathies that are associated with the aggregation and accumulation of misfolded proteins (MP). Models of MP dynamics aim at discovering the bio-mechanical propagation properties of NDs by identifying plausible associated dynamical systems. Solving these systems along the full disease course is usually challenging, due to the lack of a well defined time axis for the pathology. This issue is addressed by disease progression models (DPM) where long-term progression trajectories are estimated via time reparametrization of individual observations. However, due to their loose assumptions on the dynamics, DPM do not provide insights on the bio-mechanical properties of MP propagation. Here we propose a unified model of spatio-temporal protein dynamics based on the joint estimation of long-term MP dynamics and time reparameterization of individuals observations. The model is expressed within a Gaussian Process (GP) regression setting, where constraints on the MP dynamics are imposed through non–linear dynamical systems. We use stochastic variational inference on both GP and dynamical system parameters for scalable inference and uncertainty quantification of the trajectories. Experiments on simulated data show that our model accurately recovers prescribed rates along graph dynamics and precisely reconstructs the underlying progression. When applied to brain imaging data our model allows the bio-mechanical interpretation of amyloid deposition in Alzheimer's disease, leading to plausible simulations of MP propagation, and achieving accurate predictions of individual MP deposition in unseen data.

1 Introduction

A peculiarity of neurodegenerative diseases (NDs) is the misfolding and subsequent accumulation of pathological proteins in the brain, leading to cellular dysfunction, loss of synaptic connections, and neuronal loss [1]. Misfolded protein (MP) aggregates can self-propagate and spread the pathology between cells and tissues, along brain networks [2, 3].

A variety of mathematical models has been proposed for better insight into the kinetics governing the MP dynamical processes and for describing the effect of their propagation along brain networks [4, 5, 6, 7, 8, 9]. These MP kinetics dynamical system models are of strategic relevance, as they may provide new understanding of the mechanisms involved in NDs, and thus allow

identification of novel strategies for treatment and diagnosis. Most of these models define the propagation dynamics through diffusion equations. This modelling choice allows to reduce the number of parameters to be estimated, but comes at the expenses of an oversimplification of the dynamics governing the MP process. First, while the pathological kinetics may be assimilated to diffusive processes in short term observations, the long term evolution of NDs are unlikely to have diffusive properties. For example, the asymptotically constant behaviour of NDs may not be described by the stationary and constant rate of change specified by diffusion equations. Second, all of these models require a precise definition of the time axis to reproduce dynamics of MP propagation that are compatible with the observations, time axis which is typically not well defined in clinical data sets. To address this issue, several alternative disease progression models (DPM) have been proposed [10, 11, 12, 13, 14, 15, 16, 17]. These approaches allow to reconstruct biomarkers trajectories along the long term disease progression by "stitching" together short term individual measurements. Each subject is characterized by specific time parameters quantifying their pathological stage with respect to the estimated global group–wise evolution. However, these models provide an "apparent" description of biomarkers dynamics, without in fact elucidating the kinetics and relationships across biomarkers. This aspect indicates a limited ability of DPMs of providing a full understanding of the pathological mechanisms.

To date, no modelling framework allows for joint MP kinetics dynamical system modelling and reconstruction of the biomarkers dynamics across the whole disease long term evolution. The problem is challenging since it requires to simultaneously account for short term observations to reconstruct the long term disease progression, and to estimate group–wise dynamics parameters specified by high–dimensional dynamical systems.

In this paper we solve this problem by formulating a model for the dynamics of MP accumulation, clearance and propagation (ACP) across structural brain networks, which includes data–driven estimates of the long term protein trajectories from short term data. Figure 1 shows a schematic representation of our framework. The ACP model is formulated as a constrained regression problem in a Bayesian non–parametric setting, where the MP progression is modelled by a Gaussian Process (GP), and constraints on the MP dynamics are imposed through systems of non linear ordinary differential equations (ODEs). The Bayesian setting allows for uncertainty quantification of the MP dynamics while, to achieve tractability, the inference problem is solved via stochastic variational inference. The constrained regression framework provides a complete description of the MP dynamics, which can be subsequently used for simulating and predicting MP changes over time through forward integration of the estimated dynamical systems. The estimated MP dynamics also provide an instrument to investigate different hypotheses of MP propagation.

We test our framework against synthetic data and compare its performances in recovering the simulated evolution and the time reparameterization as compared to standard disease progression models based on monotonic constraints. Finally, we demonstrate our framework on AV45-PET data of Alzheimer's Disease (AD) subjects from the ADNI data set. We show that it allows to compare different hypothesis of MP kinetics: diffusive vs non–linear and time–varying dynamics properties (ACP). We show that the ACP model outperforms diffusive ones in terms of prediction of amyloid deposition in unseen follow-up data.

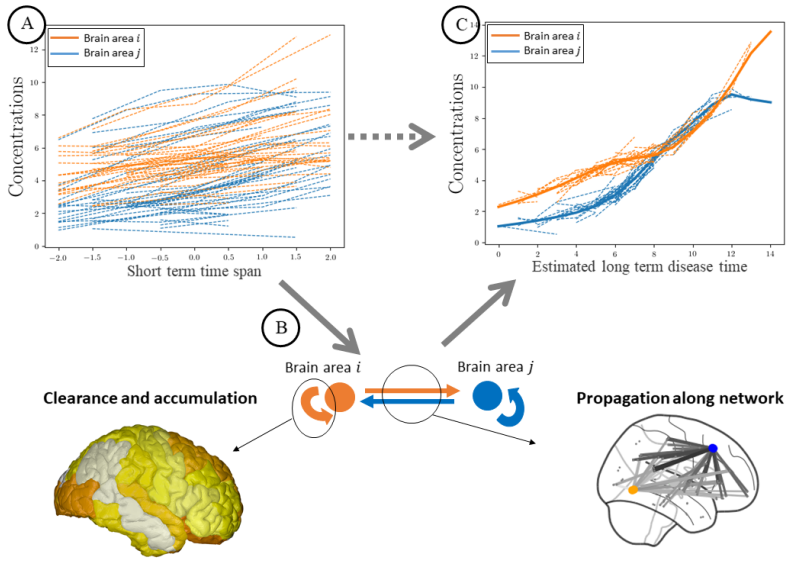


Figure 1: Schematic representation of our framework. Here we have two brain regions whose MP concentrations are collected for many subjects over a short term time span (A). The dynamics of such concentrations is described in terms of the accumulation, clearance and propagation processes, with unknown parameters (B). The proposed Bayesian framework estimates the distribution of such parameters, and the long term trajectories with respect to the estimated disease time axis (C).

2 Methods

2.1 Non–linear and time–varying MP kinetics model

We consider the brain as a system of N interconnected regions, where each region i $(i = 1, ..., N)$ is characterized by its concentration $f_i(t)$ of MP proteins along time. Standard MP kinetics models are based on the definition of dynamical systems of the form

$$\dot{\boldsymbol{f}}(t) = \beta H \boldsymbol{f}(t), \tag{1}$$

where $\boldsymbol{f}(t)$ is the vector of concentrations of MP across brain regions, H is a diffusion matrix and β is the parameter describing MP propagation. The operator H is usually defined as the graph Laplacian of the brain connectome, while β is typically assumed to be constant throughout the whole disease progression [4].

Here we introduce an extension of this paradigm which accounts for the dynamics characterized by the time–varying and non–linear parameters of MP accumulation, brain response via MP clearance, and long term propagation across neighbouring neuronal cells: the ACP model. Within this setting, equation (1) is reformulated as

$$\dot{\mathbf{f}}(t) = \mathcal{H}_{\boldsymbol{\theta}}(\boldsymbol{f}, t)\boldsymbol{f}(t), \tag{2}$$

where $\mathcal{H}_{\boldsymbol{\theta}}$ is factorized into (dependence on t is omitted but implied): $\mathcal{H}_{\boldsymbol{\theta}}(\boldsymbol{f}) = \mathcal{H}^{AC}_{\boldsymbol{\theta}}(\boldsymbol{f}) - \mathcal{H}^{out}_{\boldsymbol{\theta}}(\boldsymbol{f}) + \mathcal{H}^{in}_{\boldsymbol{\theta}}(\boldsymbol{f})$ and $\boldsymbol{\theta}$ are some kinetics parameters for accumulation, clearance and propagation. Here, $\mathcal{H}^{AC}_{\boldsymbol{\theta}}$ accounts for the total aggregation of MP plaques, i.e. sum of accumulation and clearance, while the remaining two matrices describe long–range propagation from ($\mathcal{H}^{out}_{\boldsymbol{\theta}}$) or to ($\mathcal{H}^{in}_{\boldsymbol{\theta}}$) for each region. Our assumptions on the MP dynamics are the following:

- no aggregation nor propagation occur in healthy conditions. MP plaques aggregation develops when the accumulation-clearance equilibrium breaks. This can be modelled with the assumption that $\bar{k}_a - \bar{k}_c > 0$, where $\bar{k}_\cdot$ are the maximum rates of accumulation and clearance and are assumed to be constant across regions. We define $\bar{k}_t := \bar{k}_a - \bar{k}_c$ as the maximum rate of total aggregation.

- We hypothesize a region–dependent critical threshold η_i above which the aggregation process reaches a plateau. This is modelled by a sigmoid function

$$k_t(f_i) = \frac{\bar{k_t}}{1 + e^{l_2(f_i - \eta_i)}}. \tag{3}$$

- When passing a critical threshold γ_j the MP concentration in each region j saturates and triggers propagation towards the connected regions. Also, it reaches a plateau when passing a threshold η_j. Again this can be modelled by a function asymptotically dropping to zero:

$$k_{ij}(f_j) = \frac{\bar{k}_{ij}}{\left(1 + e^{-l_1(f_j - \gamma_j)}\right)\left(1 + e^{l_2(f_j - \eta_j)}\right)}, \tag{4}$$

  representing the non–linear rate of propagation from region j to region i. Here $\bar{k}_{ij}$ is the maximum rate of propagation between the two regions, and we assume $\bar{k}_{ij} = \bar{k}_{ji}$. We combine the propagation coefficients in a matrix describing the global brain–scale propagation process: $K(\boldsymbol{f}) = (k_{ij}(f_j))_{ij}$.

- The substrate for propagation is the structural connectome, here represented by the symmetric and normalized adjacency matrix of connections between brain regions $A = (\alpha_{ij})$.

Such assumptions are formalized into the following functionals:

$$\left(\mathcal{H}^{AC}_{\boldsymbol{\theta}}(\boldsymbol{f})\right)_{ij} = \begin{cases} k_t(\boldsymbol{f}) & \text{if} \quad i = j \\ 0 & \text{otherwise;} \end{cases} \tag{5}$$

$$\mathcal{H}^{in}_{\boldsymbol{\theta}}(\boldsymbol{f}) = K(\boldsymbol{f}) \odot A; \tag{6}$$

$$\left(\mathcal{H}^{out}_{\boldsymbol{\theta}}(\boldsymbol{f})\right)_{ij} = \begin{cases} \sum_j (K(\boldsymbol{f})_{ij} \odot A_{ij}) & \text{if} \quad i = j \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Overall, the ACP model depends on $1 + 2N + \frac{N^2 - N}{2}$ parameters: $\boldsymbol{\theta} = (\bar{k_t}, \eta_i, \gamma_i, \bar{k_{ij}})$ for $i, j = 1, ..., N$.

2.2 Extending MP dynamics modelling to account for time reparametrization

Once defined our dynamical system as the one described in (2), we need to incorporate it within a regression framework for short term data. Let us assume to have S subjects for which we have measurements of MP concentrations $\mathbf{Y}$ in N brain regions at different time–points, encoded in a vector $\mathbf{t}$: $\mathbf{Y}$ is therefore the realization of $\boldsymbol{f}(t)$ at times $\mathbf{t}$. For notation simplicity we assume here $\mathbf{t}$ to be the same for every subject, but computations extend easily to more general cases.

The observations $\mathbf{Y}$ for subject k at a time points $\mathbf{t}$ can be modelled as a random sample from the following generative model [15]:

$$\mathbf{Y}^k(\mathbf{t}) = \boldsymbol{f}(\boldsymbol{\tau}^{\mathbf{k}}(\mathbf{t})) + \boldsymbol{\nu}^{\mathbf{k}} + \boldsymbol{\epsilon}. \tag{8}$$

Here $\boldsymbol{f}$ is the fixed effect function modeling the concentrations' longitudinal evolution and is modelled as a GP; $\boldsymbol{\tau}^k(\mathbf{t})$ is the individual time reparametrization with respect to the global group–wise evolution, and is modelled as a linear shift $\tau_l^k = t_l^k + d^k$ for each time point t_l^k; $\boldsymbol{\nu}^k$ is the individual random effect, assumed to be Gaussian correlated perturbations $\mathcal{N}(0, \phi_N^k)$; $\boldsymbol{\epsilon}$ is the observational noise. We introduce constraints on the dynamics of the model $\boldsymbol{f}$ enforcing the concentrations' evolution to the ACP model. This means specifying a family of admissible functions whose derivatives evaluated at the inputs $\boldsymbol{t}$ satisfy the ACP constraint:

$$\mathscr{H} = \{\boldsymbol{f}(t) : \dot{\boldsymbol{f}}(\boldsymbol{t}) = \mathcal{H}_{\boldsymbol{\theta}}(\boldsymbol{f}(\boldsymbol{t}))\boldsymbol{f}(\boldsymbol{t})\}. \tag{9}$$

We note that the constraints are imposed only on the group–wise dynamics $\boldsymbol{f}$ and not on the random–effects. This is done to reduce complexity and the model's parameters. Relaxing the constraints at individual level is also meaningful, as some subjects may be characterized by potentially different dynamics due to specific clinical conditions.

2.3 The inference scheme

We define as $\boldsymbol{F}^k$ the realization of $\boldsymbol{f}$ at $\boldsymbol{\tau}^k(\mathbf{t})$, and as $\dot{\boldsymbol{F}}^k$ the set of realizations of $\boldsymbol{f}$ and of its derivatives at $\boldsymbol{\tau}^k(\mathbf{t})$. We also indicate by $\boldsymbol{F}$, $\boldsymbol{\nu}$, and $\dot{\boldsymbol{F}}$ the collections of $\boldsymbol{F}^k$, $\boldsymbol{\nu}^k$, and $\dot{\boldsymbol{F}}^k$ for all the subjects ($k = 1, ..., S$). Similarly, we define $\boldsymbol{\tau}$ as the collections of $\boldsymbol{\tau}^k$. We denote by θ the set of parameters for the MP dynamics, and by ϕ_N the parameters associated to $\boldsymbol{\nu}$. Our framework is formulated as the constrained regression defined through two likelihood elements: a data fidelity term $p(\mathbf{Y}|\boldsymbol{F}, \phi_{\mathbf{N}}, , \mathbf{t}, \boldsymbol{\tau}, \epsilon)$ and a constraint term $p(\mathscr{H}|\dot{\boldsymbol{F}}, \theta, \mathbf{t}, \boldsymbol{\tau}, \zeta)$, where ϵ and ζ are the associated noise parameters.

Following [18] we solve the constrained regression problem by determining a lower bound for the marginal function

$$\begin{aligned} p(\mathbf{Y}, \mathscr{H}|\phi_{\mathbf{N}}, \boldsymbol{\tau}, \mathbf{t}, \epsilon, \zeta) = \int & p(\boldsymbol{C}|\boldsymbol{F}, \phi_N, \boldsymbol{\tau}, \mathbf{t}, \epsilon) p(\mathcal{A}|\dot{\boldsymbol{F}}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{t}, \zeta) \\ & p(\boldsymbol{F}, \dot{\boldsymbol{F}}|\phi_N, \boldsymbol{\tau}, \mathbf{t}) p(\boldsymbol{\theta}) d\boldsymbol{F} d\dot{\boldsymbol{F}} d\boldsymbol{\theta}, \end{aligned} \tag{10}$$

where

$$p(\boldsymbol{F}, \dot{\boldsymbol{F}}|\phi_N, \boldsymbol{\tau}, \mathbf{t}) d\boldsymbol{F} = p(\dot{\boldsymbol{F}}|\boldsymbol{F}) p(\boldsymbol{F}|\phi_N, \boldsymbol{\tau}, \mathbf{t}). \tag{11}$$

We assume the likelihood for data and constraints to be respectively Gaussian and Student–t with parameters ϵ and ζ [18], and we approximate the GP via random features expansion, as shown in [19]. Specifically, the GP realizations can be expressed as $\boldsymbol{F} \approx h(\mathbf{t}\boldsymbol{\Omega})\mathbf{W}$, where $\boldsymbol{\Omega}$ is a linear projection of the input $\mathbf{t}$ into the random feature space specified by the trigonometric activation functions $h(\cdot) = (\cos(\cdot), \sin(\cdot))$, and $\mathbf{W}$ are the regression parameters. Such approximation extends to the derivatives of the GP thanks to the chain rule [18]. As a result, the GP function and its derivatives can be both identified by the parameters $\mathbf{W}$ and $\boldsymbol{\Omega}$.

Solving (10) amounts at doing inference on $\boldsymbol{F}$, which in this setting means inference on $\mathbf{W}$ and $\boldsymbol{\Omega}$. Following [19], we optimize (10) through variational inference of $\mathbf{W}, \boldsymbol{\Omega}$ and $\boldsymbol{\theta}$. This leads to the optimization of the evidence lower bound (ELBO):

$$\begin{aligned}\log(p(\boldsymbol{C}, \mathcal{A}|\phi_N, \boldsymbol{\tau}, \boldsymbol{t}, \epsilon, \zeta)) \geq & E_{q(\boldsymbol{W})}\left[\log(p(\boldsymbol{C}|\boldsymbol{\Omega}, \boldsymbol{W}, \phi_N, \boldsymbol{\tau}, \boldsymbol{t}, \epsilon))\right] + \\ & + E_{q(\boldsymbol{W})q(\theta)}\left[\log(p(\mathcal{A}|\boldsymbol{\Omega}, \boldsymbol{W}, \theta, \boldsymbol{\tau}, \boldsymbol{t}, \zeta))\right] + \\ & - DKL(q(\boldsymbol{W})|p(\boldsymbol{W})) - DKL(q(\theta)|p(\theta)).\end{aligned} \tag{12}$$

Here $DKL(q|p)$ is the Kullback Leibler divergence between p and its variational approximation q; we assume $q(\boldsymbol{W})$ and $q(\theta)$ to be Gaussian. Details on the implementation setting are in Supplementary Material (Supp. Mat.).

3 Simulation Results

We test the ability of our framework in reconstructing the long term trajectories of the ACP dynamical system from noisy samples of short term data (Figure 2). Results are compared to the ones obtained by using the GP Progression Model [15], which includes a monotonicity constraints on the trajectories. We also test the model with data generated from a single subject and with known time–axis. Results of such simulation are in the Supp. Mat. Synthetic data are generated according to the parameters specified in Table 1. Figure 2B)-top shows the reconstructed MP trajectories from short-term data in 2A), for a two-dimensional test set. We run synthetic tests

N subjects	N regions	time interval	time–points per subject	noise
50	$\{2, 3, 11, 42\}$	$[0, 15]$	$\{1, 2, 3, 4\}$	$\mathcal{N}(0, \sigma)$, $0.2 \leq \sigma \leq 0.4$

$\bar{k}_{ij}$	$\bar{k}_t$	γ	η
$\mathcal{U}(0, 1)$	$\mathcal{U}(0, 1/2)$	$\mathcal{U}(1, \max(\frac{\boldsymbol{C}}{2}))$	$\mathcal{U}(\max(\frac{\boldsymbol{C}}{2}), \max(\boldsymbol{C}))$

Table 1: Synthetic data generation parameters.

varying the initial values of the MP parameters, the noise and the number of regions. Then, we compared our estimates of the GP and time-shift parameters with results obtained using the GP Progression Model in [15]. Table 2 shows results in terms of distributions of root mean squared errors (RMSE), for increasing number of regions. Distributions of RMSE were obtained by sampling 200 times from the estimated distributions. The ACP model generally provides better estimates for the reconstruction of the long term trajectories, as well as for the estimation of the individual time–shift as compared to the standard DPM provided by the monotonic GP. Moreover, while our framework allows the identification of the prescribed dynamical system parameters with high

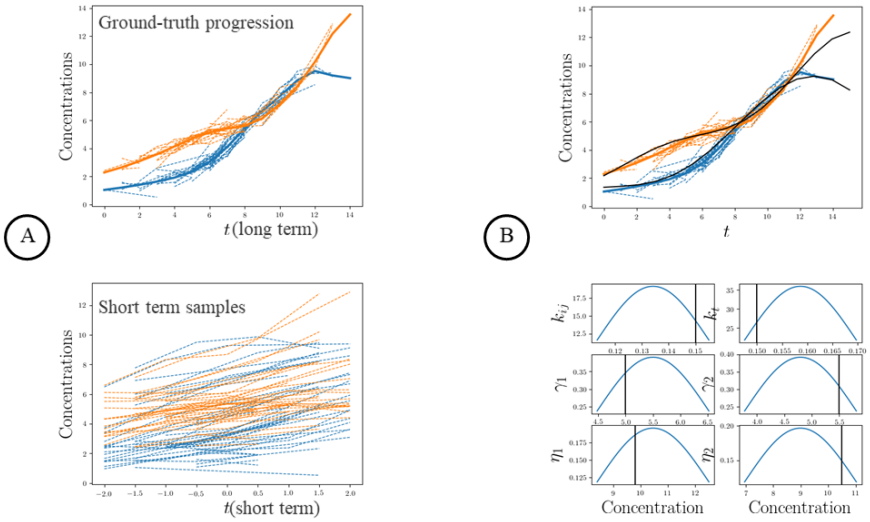


Figure 2: Results for a 2D example data. A): ground truth GP progressions and associated short term data used for benchmarking. B): ground truth and reconstructed (average) long term trajectories, and reconstructed MP parameters distributions, whose ground truth values are indicated by vertical bars.

degree of accuracy (Table 2, last row), the monotonic DPM does not allow the estimation of these quantities.

4 Modeling amyloid deposition from imaging data

4.1 Data Acquisition and Preprocessing.

4.1.1 ADNI data

This study used data from 1091 individuals from ADNI, with a total of 2380 longitudinal measurements. We collected clinical, demographic and AV45-PET SUVr data. All the subjects with either "Dementia", "Mild Cognitive Impairment" or "Cognitively Normal" clinical diagnosis were selected. We controlled for covariates (age, gender, APOE4 genotype, education) and selected 11 macro-regions, i.e. frontal, temporal, parietal, cingulate, thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accumbens, and averaged together all the values of ROIs mapped to the same macro-region, after cerebellum normalization. The macro–region definition was done to

		N=2	N=3	N=11	N=42
			Data fit		
RMSE	ACP GP	1.17(0.77)	0.93(0.37)	1.52(0.25)	1.07(0.64)
	monotonic GP	1.32(0.68)	1.08(0.52)	1.60(0.29)	1.20(0.72)
			Time–shift		
RMSE	ACP time-shift	1.67(0.49)	1.92(0.64)	1.53(0.42)	1.19(0.39)
	monotonic time–shift	1.87(0.44)	1.97(0.58)	1.62(0.42)	1.20(0.41)
			Dynamical parameters		
relative	ACP GP	6.3%	9.6%	11.4%	21.9%
error	monotonic GP	–	–	–	–

Table 2: RMSE results for GP fit, time–shifts estimates and dynamical parameters for both the GP Progression Model with monotonicity (GP) and the ACP model. The error for the dynamical parameters is expressed, in percentage, relatively to the ground truth parameters.

reduce computational expenses and aid interpretability of the resulting MP parameters. Demographic and clinical details are in Supp. Mat. We split the data set in two parts: the D1 data set contains all the longitudinal data for each subject up to the second-to-last time points. The remaining time–points were included in a second data set D2. Subjects with one measurement only were included in D1. Data set D1 includes 1651 longitudinal data of 1091 subjects; D2 contains 731 cross-sectional data. We run the models on D1, estimating MP dynamics, GP parameters and individual time–shifts, and used D2 to validate model predictions.

4.1.2 HCP data.

Data used in the preparation of this work were obtained from the MGH-USC Human Connectome Project database. We collected 3D T1w and DTI of 24 age and gender–matched subjects. The pipeline for structural connectome generation is described in [22]. We averaged the 24 connectomes together and obtained an average young, healthy connectome. Finally, we averaged together the regions belonging to the same lobe or subcortical area (to obtain 11 macro–region) and we set to 0 all the weights below the average weights across nodes, and to 1 the weight above. This last step was performed in order to remove the weak connections.

4.2 Estimated long term dynamics.

We analyzed the AV45–PET data with two different models of MP kinetics: the ACP model of equation (2) - which has non–linear and time–varying dynamics, and a full diffusive model. The diffusion dynamics were prescribed by the system $\dot{\boldsymbol{c}}(t) = B\boldsymbol{c}(t)$, where the coefficients b_{ij} of B are estimated (along with individual time parameters) with our framework. Figure 3A) shows the long term trajectories estimated with both models, for four regions. Figure 3B) shows the time associated to each regional trajectory at which maximal separation between "Cognitively Normal" and "Alzheimer's disease" subjects was measured. The time distribution is inferred from the trajectory distribution associated to each region. The dynamics and orderings of ACP and diffusion provide plausible description of the pathological evolution of amyloid deposition, compatible with previous findings in histo pathological and imaging studies in AD [20, 21].

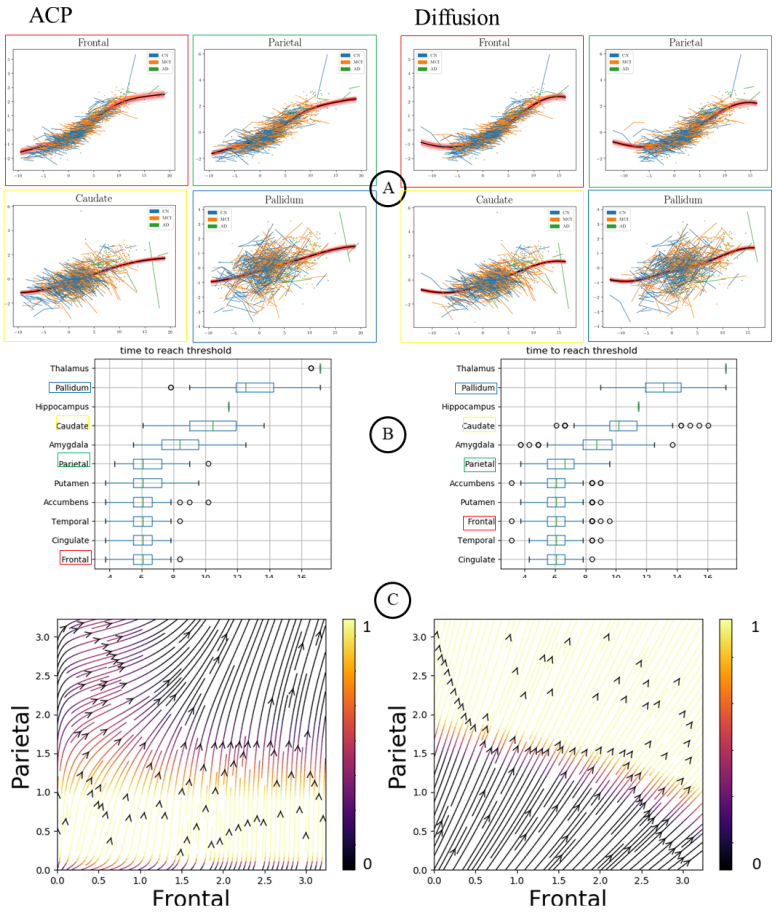


Figure 3: A): estimated long term trajectories and individual short term measurements for 4 regions of interest: frontal lobe, parietal lobe, caudate and pallidum, for the two models. B): ordering derived from the trajectories. Regions visualized in A) are highlighted. C): streamlines of the 2D fields in the {Frontal, Parietal} plane.

4.3 Predictions performances of the models.

Figure 3C) shows the estimated vector fields for the relative dynamics of the frontal and parietal lobes. This vector field is obtained by integrating the dynamical system estimated for respectively ACP and diffusion models. Therefore it does not correspond to extrapolation of the curves in Figure 3A). Here the other biomarkers are set constant to their mean values. We can appreciate the non–linear dynamics of the ACP model, as well as the linear dynamics of the diffusive model. The resulting vector fields provide a tool for interpreting and comparing mechanistic hypotheses. Indeed, Figure 3C) shows that the ACP model estimates an initial fast propagation, which slows down with time. The opposite behaviour is observed by analyzing the dynamics of the diffusion model, with an acceleration in the propagation along with the progression. This behaviour is

unlikely to reproduce real–case scenarios, where amyloid aggregation eventually slows down and does not accumulates indefinitely. This result points to the higher biological plausibility of the proposed ACP model. For each subject in D1 with follow-up measurements in D2, we computed the streamline associated with their individual dynamics (in the whole 11-D space), and estimated the values of each biomarker at the corresponding follow-up time. We computed the RMSE for each estimate, and bootstrapped over the MP dynamic parameters 200 times, obtaining RMSE distributions (Table 3).

	frontal	temporal	parietal	cingulate
ACP	0.21(0.16)	0.18(0.14)	0.20(0.16)	0.20(0.15)
diffusion	0.25(0.18)	0.25(0.17)	0.22(0.19)	0.24(0.18)

	thalamus	caudate	putamen	pallidum	hippo	amygdala	accumbens
ACP	0.12(0.09)	0.16(0.12)	0.16(0.13)	0.13(0.10)	0.12(0.09)	0.12(0.10)	0.21(0.15)
diffusion	0.11(0.08)	0.17(0.13)	0.17(0.13)	0.17(0.13)	0.12(0.09)	0.11(0.09)	0.24(0.19)

Table 3: RMSE (mean, sd) for the ACP and the diffusion models estimates. The ACP model generally provides predictions closer to the observed follow-up values.

4.4 Misfolded proteins propagation pathways.

Figure 4 shows the connectomes where the edges' colors are set to be proportional to the values of the estimated MP parameters for the ACP model (plot on the left hemisphere), or to the values of the propagation parameters for the diffusive model (plot on the right hemisphere). The parameters have been normalized to $[0, 1]$ to aid comparison. The paths appear to be different for the two models and the ACP model seems to better describe the frontal–posterior pathway known to characterize amyloid deposition in AD [20, 21].

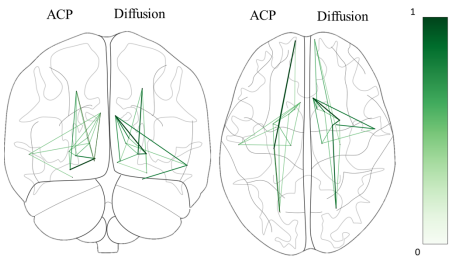


Figure 4: Coronal and axial views of connectomes with edges' colors proportional to the values of the estimated propagation parameters for either the ACP model (left hemisphere) and the diffusion model (right hemisphere).

5 Discussion

We presented a spatio–temporal model of MP dynamics over brain networks. The model is based on the joint estimation of long term MP dynamics and time reparametrization of individuals observations, and is expressed within a GP regression setting, where constraints on the MP dynamics are imposed through non–linear dynamical systems, which account for accumulation, clearance and propagation of MP. Experiments on simulated data show that our model accurately recovers prescribed rates along graph dynamics and precisely reconstructs the underlying progression. When applied to AV45-PET brain imaging data our model allows the bio-mechanical interpretation of amyloid deposition in Alzheimer's disease, leading to plausible simulations of MP propagation, and achieving accurate predictions of individual MP deposition in unseen data.

The method has some limitations: first of all, structural connectome estimation using tractography is known to be prone to false positive and negative connections. Nevertheless, here we take an average connectome over multiple young and healthy subjects, which we believe works a reasonable anatomical reference. Another limitation of the model is that it assumes that all subjects follow the same disease progression pattern, which might not be the case in heterogeneous data sets such as ADNI.

The ideas we propose here extend to a much larger range of diseases and alternative models of propagation, such as propagation via functional networks [5, 9], or different kind of tractography to represent intra– and extra–axonal propagation [22].

References

[1] Soto, C. and Pritzkow, S., 2018. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. Nat Neurosci, 21(10), pp.1332-1340.

[2] Jucker, M. and Walker, L.C., 2013. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. Nature, 501(7465), p.45.

[3] Brettschneider, J., Del Tredici, K., Lee, V.M.Y. and Trojanowski, J.Q., 2015. Spreading of pathology in neurodegenerative diseases: a focus on human studies. Nature Reviews Neuroscience, 16(2), p.109.

[4] Raj, A., Kuceyeski, A. and Weiner, M., 2012. A network diffusion model of disease progression in dementia. Neuron, 73(6), pp.1204-1215.

[5] Zhou, J., Gennatas, E.D., Kramer, J.H., Miller, B.L. and Seeley, W.W., 2012. Predicting regional neurodegeneration from the healthy brain functional connectome. Neuron, 73(6), pp.1216-1227.

[6] Iturria-Medina, Y., Sotero, R.C., Toussaint, P.J., Evans, A.C. and Alzheimer's Disease Neuroimaging Initiative, 2014. Epidemic spreading model to characterize misfolded proteins propagation in aging and associated neurodegenerative disorders. PLoS computational biology, 10(11), p.e1003956.

[7] Raj, A., LoCastro, E., Kuceyeski, A., Tosun, D., Relkin, N., Weiner, M. and Alzheimer's Disease Neuroimaging Initiative (ADNI, 2015. Network diffusion model of progression predicts

longitudinal patterns of atrophy and metabolism in Alzheimer's disease. Cell reports, 10(3), pp.359-369.

[8] Weickenmeier, J., Kuhl, E. and Goriely, A., 2018. Multiphysics of Prionlike Diseases: Progression and Atrophy. Physical review letters, 121(15), p.158101.

[9] Cauda, F., Nani, A., Manuello, J., Premi, E., Palermo, S., Tatu, K., Duca, S., Fox, P.T. and Costa, T., 2018. Brain structural alterations are distributed following functional, anatomic and genetic connectivity. Brain, 141(11), pp.3211-3232.

[10] Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C. and Alexander, D.C., 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage, 60(3), pp.1880-1889.

[11] Villemagne, V.L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K.A., Salvado, O., Szoeke, C., Macaulay, S.L., Martins, R., Maruff, P. and Ames, D., 2013. Amyloid beta deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. The Lancet Neurology, 12(4), pp.357-367.

[12] Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M. and Alexander, D.C., 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain, 137(9), pp.2564-2577.

[13] Donohue, M.C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R.G., Raman, R., Gamst, A.C., Beckett, L.A., Jack Jr, C.R., Weiner, M.W., Dartigues, J.F. and Aisen, P.S., 2014. Estimating long-term multivariate progression from short-term data. Alzheimer's and Dementia, 10(5), pp.S400-S410.

[14] Schiratti, J.B., Allassonniere, S., Colliot, O. and Durrleman, S., 2015. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In Advances in Neural Information Processing Systems (pp. 2404-2412).

[15] Lorenzi, M., Filippone, M., Frisoni, G.B., Alexander, D.C., Ourselin, S. and Alzheimer's Disease Neuroimaging Initiative, 2017. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. Neuroimage.

[16] Marinescu, R.V., Eshaghi, A., Lorenzi, M., Young, A.L., Oxtoby, N.P., Garbarino, S., Crutch, S.J., Alexander, D.C. and Alzheimer's Disease Neuroimaging Initiative, 2019. DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. NeuroImage.

[17] Koval, I., Schiratti, J.B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., Durrleman, S. and Alzheimer's Disease Neuroimaging Initiative, 2017, September. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 451-459). Springer, Cham.

[18] Lorenzi, M. and Filippone, M., 2018. Constraining the Dynamics of Deep Probabilistic Models. Proceedings of the 35th International Conference on Machine Learning, 80, pp.3233-3242.

[19] Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M., 2017. Random feature expansions for deep Gaussian processes. Proceedings of the 34th International Conference on Machine Learning, 70, pp.884-893.

[20] Thal, D.R., Rub, U., Orantes, M. and Braak, H., 2002. Phases of Abeta-deposition in the human brain and its relevance for the development of AD. Neurology, 58(12), pp.1791-1800.

[21] Irvine, G.B., El-Agnaf, O.M., Shankar, G.M. and Walsh, D.M., 2008. Protein aggregation in the brain: the molecular basis for Alzheimer's and Parkinson's diseases. Molecular medicine, 14(7-8), pp.451-464.

[22] Oxtoby, N.P., Garbarino, S., Firth, N.C., Warren, J.D., Schott, J.M., Alexander, D.C. and Alzheimer's Disease Neuroimaging Initiative, 2017. Data-Driven sequence of changes to anatomical Brain connectivity in sporadic Alzheimer's Disease. Frontiers in neurology, 8, p.580.

Monotonic Gaussian Process for Spatio-Temporal Disease Progression Modeling in Brain Imaging Data

Clément Abi Nader[1], Nicholas Ayache[1], Philippe Robert[2] and Marco Lorenzi[1]

1. Université Côte d'Azur, INRIA Sophia Antipolis, EPIONE research group, France

2. Université Côte d'Azur, CoBTeK lab, MNC3 program, France

Originally published in:
NeuroImage, 2019, to appear

Abstract

We introduce a probabilistic generative model for disentangling spatio-temporal disease trajectories from collections of high-dimensional brain images. The model is based on spatio-temporal matrix factorization, where inference on the sources is constrained by anatomically plausible statistical priors. To model realistic trajectories, the temporal sources are defined as monotonic and time-reparameterized Gaussian Processes. To account for the non-stationarity of brain images, we model the spatial sources as sparse codes convolved at multiple scales. The method was tested on synthetic data favourably comparing with standard blind source separation approaches. The application on large-scale imaging data from a clinical study allows to disentangle differential temporal progression patterns mapping brain regions key to neurodegeneration, while revealing a disease-specific time scale associated to the clinical diagnosis.

1 Introduction

Neurodegenerative disorders such as Alzheimer's disease (AD) are characterized by morphological and molecular changes of the brain, ultimately leading to cognitive and behavioral decline. Clinicians suggested hypothetical models of the disease evolution, showing how different types of biomarkers interact and lead to the final dementia stage [15]. In the past years, efforts have been made in order to collect large databases of imaging and clinical measures, hoping to obtain more insights about the disease progression through data-driven models describing the trajectory of the disease over time. This kind of models are of critical importance for understanding the pathological progression in large scale data, and would represent a valuable reference for improving the individual diagnosis.

Current clinical trials in AD are based on longitudinal monitoring of biomarkers. Disease progression modelling aims at providing an interpretable way of modelling the evolution of biomarkers according to an estimated history of the pathology, as proposed for example in [10], [12], [16], [25], and [41]. Therefore, disease progression models are promising methods for automatically staging patients, and quantifying their progression with respect to the underlying model of the pathology. These approaches entail a great potential for automatic stratification of individuals

based on their estimated stage and progression speed, and for assessment of efficacy of disease modifying drugs. Within this context, we propose a spatio-temporal generative model of disease progression, aimed at disentangling and quantifying the independent dynamics of changes observed in datasets of multi-modal data. With this term we indicate data acquired via different imaging modalities such as Magnetic Resonance Imaging (MRI) or Positron-Emission Tomography (PET), as well as non-imaging data such as clinical scores assessed by physicians. Moreover, we aim at automatically inferring the disease severity of a patient with respect to the estimated trajectory. Defining such a disease progression model raises a number of methodological challenges.

AD spreads over decades with a temporal mismatch between the onset of the disease and the moment where the clinical symptoms appear. Either age of diagnosis, or the chronological age, are therefore not suitable as a temporal reference to describe the disease progression in time. Moreover, as the follow-up of patients doesn't exceed a few years, the development of a model of long-term pathological changes requires to integrate cross-sectional data from different individuals, in order to consider a longer period of time. In virtue of the lack of a well defined temporal reference, observations from different individuals are characterized by large and unknown variability in the onset and speed of the disease. It is therefore necessary to account for a time-reparameterization function, mapping each individuals' observations to a common temporal axis associated to the absolute disease trajectory [16, 36]. This would allow to estimate an absolute time-reference related to the natural history of the pathology.

The analysis of MRI and PET data, requires to account for spatio-temporally correlated features (voxels, i.e. volumetric pixels) defined over arrays of more than a million entries. The development of inference schemes jointly considering these correlation properties thus raises scalability issues, especially when accounting for the non-stationarity of the image signal. Furthermore, the brain regions involved in AD exhibit various dynamics in time, and evolve at different speed [39]. From a modeling perspective, accounting for differential trajectories over space and time raises the problem of source identification and separation. This issue has been widely addressed in neuroimaging via Independent Component Analysis (ICA) [8], especially on functional MRI (fMRI) data [7]. Nevertheless, while fMRI time-series are usually defined over a few hundreds of time points acquired per subject, our problem consists in jointly analyzing short-term and cross-sectional data observations with respect to an unknown time-line. This problem cannot be tackled with standard ICA, as time is generally an independent variable on which inference is not required. Moreover, ICA retrieves spatial sources based on the assumption of statistical independence. This assumption does not necessarily lead to clinically interpretable findings. Indeed, dependency across temporal patterns can be still highly relevant to the pathology, for example when modeling temporal delay across similar sources.

The problem of providing a realistic description of the biological processes is critical when analyzing biomedical data, such as medical images. For example, to describe a plausible evolution of AD from normal to pathological stages, smoothness and monotonicity are commonly assumed for the temporal sources. It is also necessary to account for the non-stationarity of changes affecting the brain from global to localized spatio-temporal processes. As a result, spatial sources need to account for different resolutions at which these changes take place. While several multi-scale analysis approaches have been proposed to model spatio-temporal signals [26, 6, 14], extending this type of methods to the high-dimension of medical images is generally not trivial due to scalability

issues. Finally, the noisy nature of medical images, along with the large signal variability across observations, requires a modeling framework robust to bias and noise.

In this work, we propose to jointly address these issues within a Bayesian framework for the spatio-temporal analysis of large-scale collections of multi-modal brain data. We show that this framework allows us to naturally encode plausibility constraints through clinically-inspired priors, while accounting for the uncertainty of the temporal profiles and brain structures we wish to estimate. Similarly to the ICA setting, we formulate the problem of trajectory modeling through matrix factorization across temporal and spatial sources. This is done for each modality by inferring their specific spatio-temporal sources. To promote smoothness in time and avoid any unnecessary hypothesis on the temporal trajectories, we rely on non-parametric modeling based on Gaussian Process (GP). We account for a plausible evolution from healthy to pathological stages thanks to a monotonicity constraint applied on the GP. Moreover, individuals' observations are temporally re-aligned on a common scale via a time-warping function. In case of imaging data, to model the non-stationarity of the spatial signal, the spatial sources are defined as sparse activation maps convolved at different scales. We show that our framework can be efficiently optimized through stochastic variational inference, allowing to exploit automatic differentiation and GPU support to speed up computations.

The paper is organized as follows: Section 2 analyzes related work on spatio-temporal modeling of neurodegeneration, while Section 3 details our method. In Section 4 we present experiments on synthetic data in which we compare our model to standard blind source separation approaches. We finally provide a demonstration of our method on the modeling of imaging data from a large scale clinical study. Prospects for future work and conclusions are drawn in section 5. Derivations that we could not fit in the paper are detailed in the Appendices.

2 Related Work in Neurodegeneration Modeling

To deal with the uncertainty of the time-line of neurodegenerative pathologies, the concept of time-reparameterization of imaging-derived features has been used in several works. The underlying principle consists in estimating an absolute time-scale of disease progression by temporally re-aligning data from different subjects. For instance, in [42] the time-evolution was approximated as a sequence of events which need to be re-ordered for each patient. This approach thus considers the evolution of neurodegenerative diseases as a collection of transitions between discrete stages. This hypothesis is however limiting, as it doesn't reflect the continuity of changes affecting the brain along the course of the pathology.

To address this limitation, we rely on a continuous parameterization of the time-axis as in [25, 10]. In particular, individuals' observations are time-realigned on a common temporal scale via a time-warping function. Using a set of relevant scalar biomarkers, this kind of approach allows to learn a time-scale describing the pathology evolution, and to estimate a data-driven time-line markedly correlated with the decline of cognitive abilities. Similarly, in [4] a disease progression score was estimated using biomarkers from molecular imaging. These methods are however based on the analysis of low-dimensional measures, such as collections of clinical variables. Therefore, they do not allow to scale to the high dimension of multi-modal medical images. Our work tackles

this shortcoming thanks to a scalable inference scheme based on stochastic variational inference.

Concerning the spatio-temporal representation of neurodegeneration, a mixed-effect model was proposed by [21] to learn an average spatio-temporal trajectory of brain evolution on cortical thickness data. The fixed-effect describes the average trajectory, while random effects are estimated through individual spatio-temporal warping functions, modeling how each subject differs from the global progression. Still, the extension of this approach to image volumes raises scalability issues. It has also to be noted that, to allow computational tractability, the brain evolution was assumed to be stationary both in space and time, thus limiting the ability of the model to disentangle the multiple dynamics of the brain structures involved in AD.

An attempt to source separation is proposed in [27], through the decomposition of cortical thickness measurements as a mixture of spatio-temporal processes. This is performed by associating to each cortical vertex a temporal progression modeled by a sigmoid function, which may be however too simplistic to describe the progression of AD temporal processes. We propose to overcome this issue by non-parametric modeling of the temporal sources through GPs. Moreover, the model in [27] is lacking of an explicit vertex-wise correlation model, as it only assumes correlation between clustering parameters at the resolution of the mesh graph. For this reason, it may still be sensitive to spatial variation at different scales and noise. We address this problem by modeling the spatial sources through convolution of sparse maps at multiple resolutions, allowing to deal with signal non-stationarity and robustness to noise.

3 Methods

In the following sections a matrix will be denoted by an uppercase letter $\boldsymbol{X}$, its n-th row will be given by $\boldsymbol{X}_{n:}$ and its n-th column by $\boldsymbol{X}_{:n}$. A column vector will be denoted by a lowercase letter $\boldsymbol{x}$. Subscript indices will be used to index the elements of matrices, vectors or sets of scalars. Superscipt indices will allow to index the blocks of block diagonal matrices.

3.1 Individual time-shift

To account for the uncertainty of the time-line of individual measurements, we assume that the observations are defined with respect to an absolute temporal reference τ. This is performed through a time-warping function $t_p = \boldsymbol{f}_p(\tau)$, that models the individual time-reparameterization. We choose an additive parameterization such that:

$$\boldsymbol{f}_p(\tau) = \tau + \delta_p. \tag{1}$$

Within this setting the individual time-shift δ_p encodes the temporal position of subject p, which in our application can be interpreted as the disease stage of subject p with respect to the long-term disease trajectory. We denote by $\boldsymbol{\delta} = \{\delta_p\}_{p=0}^{P}$ the set of time-shift parameters.

3.2 Data modeling

We represent the spatio-temporal data $\boldsymbol{D}$ by a block diagonal matrix in which we differentiate two main blocks $\boldsymbol{Y}$ and $\boldsymbol{V}$ as illustrated in Figure 1. Each sub-block $\boldsymbol{Y}^m$ is a matrix containing

the data represented by one of the M imaging modalities we wish to consider. These matrices have dimensions $P \times F_m$, where P denotes the number of subjects and F_m the number of imaging features for modality m, which in our case is the number of voxels. The matrix $\boldsymbol{V}$ accounts for non-imaging or scalar data such as clinical scores and has dimensions $P \times C$, where C is the number of scalar features considered. We postulate a generative model and decompose the data as shown in Figure 1. For each sub-block $\boldsymbol{Y}^m$, the data is factorized in a set of N_m spatio-temporal sources

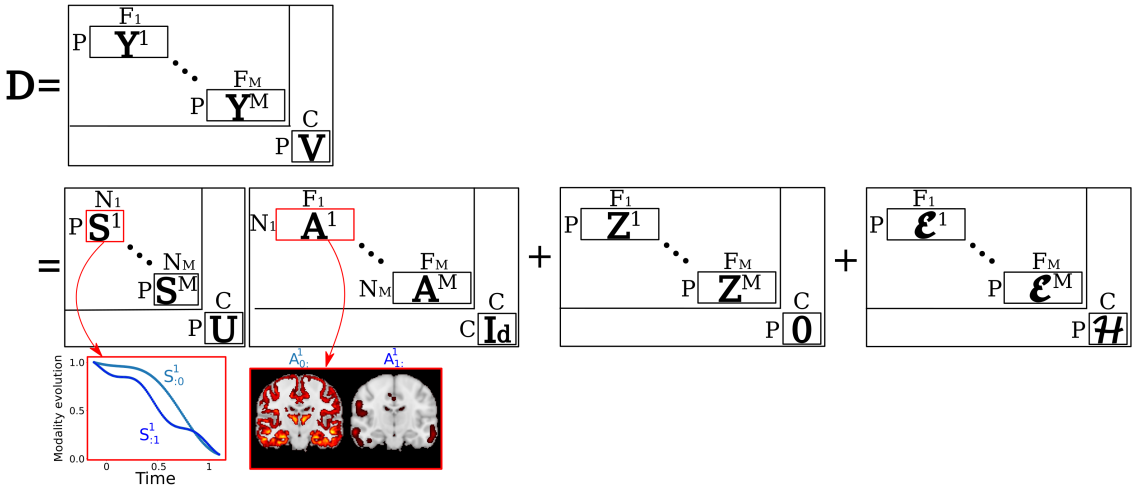


Figure 1: Spatio-temporal decomposition of each data block. A data matrix composed by M imaging modalities is decomposed as the product of monotonic temporal sources $\boldsymbol{S}^m$ and corresponding activation maps $\boldsymbol{A}^m$. Monotonic sources are also used to model the scalar biomarkers $\boldsymbol{V}$, while we assume additive constant terms $\boldsymbol{Z}^m$, and noise $\boldsymbol{\mathcal{E}}^m$.

$\boldsymbol{Y}^m = \boldsymbol{S}^m \boldsymbol{A}^m$. The columns of the matrix $\boldsymbol{S}^m$ describe the non-linear temporal evolution of the corresponding spatial maps contained in the rows of $\boldsymbol{A}^m$. Therefore, their product represents the voxel-wise linear combination of the spatial maps modulated by the corresponding temporal sources. The subjects share the same set of temporal sources across $\boldsymbol{S}^1, .., \boldsymbol{S}^M$, as these sources describe the temporal evolution of the group-wise images through the regression problem specified in Figure 1. The data in matrix $\boldsymbol{V}$ is modelled by a matrix $\boldsymbol{U}$ whose columns depict the temporal trajectories of the different scalar scores. In the case of imaging data, we also consider a constant term modeling brain areas which don't exhibit any intensity changes over time. This is done by including constant matrix terms $\boldsymbol{Z}^m$ that we need to estimate. We assume for a given modality m that the vectors $\boldsymbol{Z}^m_{p:}$ are common to every subjects. Finally, for each modality m, scalar score c, and subject p, we assume Gaussian observational noise $\boldsymbol{\mathcal{E}}^m_{p:} \sim \mathcal{N}(\boldsymbol{0}, \sigma_m^2 \boldsymbol{I})$, and $\boldsymbol{\mathcal{H}}_{p,c} \sim \mathcal{N}(0, \nu_c^2)$ for respectively imaging and scalar information.

Therefore, if we consider the data from modality m and scalar c of patient p observed at time $\boldsymbol{f}_p(\tau)$ we have:

$$\begin{aligned} \boldsymbol{Y}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m, \psi_m) &= \boldsymbol{S}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m) \boldsymbol{A}^m(\psi_m) + \boldsymbol{Z}^m_{p:} + \boldsymbol{\mathcal{E}}^m_{p:}, \\ \boldsymbol{V}_{p,c}(\boldsymbol{f}_p(\tau), \theta_c) &= \boldsymbol{U}_{p,c}(\boldsymbol{f}_p(\tau), \theta_c) + \boldsymbol{\mathcal{H}}_{p,c}. \end{aligned} \tag{2}$$

We denote by θ_m and θ_c the temporal parameters related respectively to the modality m and scalar feature c, while ψ_m represents the set of spatial parameters of modality m. We assume conditional independence across modalities and scalar scores given the time-shift information:

$$p(\boldsymbol{Y}, \boldsymbol{V}|\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{\delta}, \sigma, \nu) = \Big(\prod_m p(\boldsymbol{Y}^m|\boldsymbol{A}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) \Big) \Big(\prod_c p(\boldsymbol{V}_{:c}|\boldsymbol{U}_{:c}, \boldsymbol{\delta}, \nu_c) \Big). \quad (3)$$

Relying on classical regression formulation, we assume exchangeability across subjects allowing us to derive the data likelihood for a given modality m. According to the generative model we can write:

$$\begin{aligned} p(\boldsymbol{Y}^m|\boldsymbol{A}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) = \Big(\prod_p \frac{1}{(2\pi\sigma_m^2)^{\frac{F_m}{2}}} \exp(-\frac{1}{2\sigma_m^2} ||\boldsymbol{Y}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m, \psi_m) \\ - \boldsymbol{S}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m)\boldsymbol{A}^m(\psi_m) - \boldsymbol{Z}^m_{p:}||^2) \Big). \end{aligned} \quad (4)$$

Naturally, a similar equation holds for $p(\boldsymbol{V}_{:c}|\boldsymbol{U}_{:c}, \boldsymbol{\delta}, \nu_c)$.

Within a Bayesian modeling framework, we wish to maximize the marginal log-likelihood $\log(p(\boldsymbol{Y}, \boldsymbol{V}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu))$, to obtain posterior distributions for the spatio-temporal processes. Since the derivation of this quantity in a closed-form is not possible, we tackle this optimization problem through stochastic variational inference. Based on this formulation, in what follows we illustrate our model by detailing the variational approximations imposed on the spatio-temporal sources, along with the priors and constraints we impose to represent the data (Sections 3.3 and 3.4). Finally, we detail the variational lower bound and optimization strategy in Section 3.5.

For ease of notation we will drop the m and c indexes in Sections 3.3 and 3.4. As a result the matrix $\boldsymbol{S}$ will indistinctly refer to either any $\boldsymbol{S}^m$ or $\boldsymbol{U}$, while matrix $\boldsymbol{A}$ will refer to any $\boldsymbol{A}^m$, and $\boldsymbol{Y}$ to any $\boldsymbol{Y}^m$. For a given modality m, the number of patients P will be indexed by p, the number of sources N^m or the number of scalar scores C will be indexed by n, and finally f will index the number of imaging features F^m.

3.3 Spatio-temporal processes

3.3.1 Temporal sources

In order to flexibly account for non-linear temporal patterns, the temporal sources are encoded in a matrix $\boldsymbol{S}$ in which each column $\boldsymbol{S}_{:n}$ is a GP representing the evolution of source n and is independent from the other sources. To allow computational tractability within a variational setting, we rely on the GP approximation proposed in [9], through kernel approximation via random feature expansion [32]. Within this framework, a GP can be approximated as a Bayesian Neural Network with form: $\boldsymbol{S}_{:n}(\boldsymbol{t}) = \phi(\boldsymbol{t}(\boldsymbol{\omega}^n)^T)\boldsymbol{w}^n$. For example, in the case of the Radial Basis Function (RBF) covariance, $\boldsymbol{\omega}^n$ is a linear projection in the spectral domain. It is equipped with a Gaussian distributed prior $p(\boldsymbol{\omega}^n) \sim \mathcal{N}(\boldsymbol{0}, l_n\boldsymbol{I})$ with a zero-mean and a covariance parameterized by a scalar l_n, acting as the length-scale parameter of the RBF covariance. The non-linear basis functions activation is defined by setting $\phi(\cdot) = (\cos(\cdot), \sin(\cdot))$, while the regression parameter $\boldsymbol{w}^n$ is given with a standard normal prior. The GP inference problem can be conveniently performed by estimating approximated variational distributions for all the $\boldsymbol{\omega}^n$ and $\boldsymbol{w}^n$ (Section 3.5). We will respectively denote by $\boldsymbol{\Omega}$ and

$\boldsymbol{W}$ the block diagonal matrices whose blocks are the $(\boldsymbol{\omega}^n)^T$ and $\boldsymbol{w}^n$. Considering the N temporal sources, we can write $p(\boldsymbol{\Omega}) = \prod_n p(\boldsymbol{\omega}^n)$ and $p(\boldsymbol{W}) = \prod_n p(\boldsymbol{w}^n)$.

We wish also to account for a steady evolution of the temporal processes, hence constraining the temporal sources to monotonicity. This is relevant in the medical case, where one would like to model the steady progression of a disease from normal to pathological stages. In our case, we want to constrain the space of the temporal sources to the set of solutions $\mathcal{C}_n = \{\boldsymbol{S}_{:n}(\boldsymbol{t}) \mid \boldsymbol{S}'_{:n}(\boldsymbol{t}) \geq 0 \quad \forall\, \boldsymbol{t}\}$. This can be done done consistently within the regression setting of [33], and in particular with the GP random feature expansion framework as shown in [24]. In that work, the constraint is introduced as a second likelihood term on the temporal sources dynamics:

$$p(\mathcal{C}|\boldsymbol{S}', \gamma) = \prod_{p,n} (1 + \exp(-\gamma \boldsymbol{S}'_{p,n}(\boldsymbol{t})))^{-1}, \tag{5}$$

where $\boldsymbol{S}'$ contains every derivatives $\boldsymbol{S}'_{:n}$, γ controls the magnitude of the monotonicity constraint, and $\mathcal{C} = \bigcap_n \mathcal{C}_n$. According to [24] this constraint can be specified through the parametric form for the derivative of each $\boldsymbol{S}_{:n}$:

$$\boldsymbol{S}'_{:n}(t) = \frac{d\phi(\boldsymbol{t}(\boldsymbol{\omega}^n)^T)}{d\boldsymbol{t}} \boldsymbol{w}^n. \tag{6}$$

This setting leads to an efficient scheme for estimating the temporal sources through stochastic variational inference (Section 3.5).

3.3.2 Spatial sources.

According to the model introduced in Section 3.2, each observation $\boldsymbol{Y}_{p:}$ is obtained as the linear combination at a specific time-point between the temporal and spatial sources. In order to deal with the multi-scale nature of the imaging signal, we propose to represent the spatial sources at multiple resolutions. To this end, we encode the spatial sources in a matrix $\boldsymbol{A}$ whose rows $\boldsymbol{A}_{n:}$ represent a specific source at a given scale. The scale is prescribed by a convolution operator $\boldsymbol{\Sigma}^n$, which is a applied to a map $\boldsymbol{B}_{n:}$ that we wish to infer. This problem can be specified by defining $\boldsymbol{A}_{n:} = \boldsymbol{B}_{n:}\boldsymbol{\Sigma}^n$, where $\boldsymbol{\Sigma}^n$ is an $F \times F$ Gaussian kernel matrix imposing a specific spatial resolution. The length-scale parameter λ_n of the Gaussian kernel is fixed for each source, to force the model to pick details at that specific scale. Due to the high-dimension of the data we are modeling, performing stochastic variational inference in this setting raises scalability issues. For instance, if we assume a Gaussian distribution $\mathcal{N}(\mu_{\boldsymbol{B}_{n:}}, diag(\boldsymbol{\Lambda}))$ for $\boldsymbol{B}_{n:}$, the distribution of the spatial signal would be $p(\boldsymbol{A}_{n:}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{B}_{n:}}\boldsymbol{\Sigma}^n, \boldsymbol{\Sigma}^n diag(\boldsymbol{\Lambda})(\boldsymbol{\Sigma}^n)^T)$. As a result, sampling from $p(\boldsymbol{A}_{n:})$ is not computationally tractable due to the size of the covariance matrix, which prevents the use of standard inference schemes on $\boldsymbol{B}_{n:}$. This can be overcome thanks to the separability of the Gaussian convolution kernel [28, 23], according to which the 3D convolution matrix $\boldsymbol{\Sigma}^n$ can be decomposed into the Kronecker product of 1D matrices, $\boldsymbol{\Sigma}^n = \boldsymbol{\Sigma}^n_x \otimes \boldsymbol{\Sigma}^n_y \otimes \boldsymbol{\Sigma}^n_z$. This decomposition allows to efficiently perform standard operations such as matrix inversion, or matrix-vector multiplication [35]. Thanks to this choice, we recover tractability for the inference of $\boldsymbol{B}_{n:}$ through sampling, as required by stochastic inference methods [20].

3.4 Sparsity

In order to detect specific brain areas involved in neurodegeneration, we propose to introduce a sparsity constraint on the maps (or codes) $\boldsymbol{B}_{n:}$. Consistently with our variational inference scheme, we induce sparsity via *Variational Dropout* as proposed in [19]. This approach leverages on an improper log-scale uniform prior $p(|\boldsymbol{B}_{n:}|) \propto \prod_f \ 1/|\boldsymbol{B}_{n,f}|$, along with an approximate posterior distribution:

$$q_1(\boldsymbol{B}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{M}_{n:}, diag(\alpha_{n,1}\boldsymbol{M}_{n,1}^2 ... \alpha_{n,F}\boldsymbol{M}_{n,F}^2)). \tag{7}$$

In this formulation, the dropout parameter $\alpha_{n,f}$ is related to the individual dropout probability $p_{n,f}$ of each weight by $\alpha_{n,f} = p_{n,f}(1-p_{n,f})^{-1}$. When the parameter $\alpha_{n,f}$ exceeds a fixed threshold, the dropout probability $p_{n,f}$ is considered high enough to ignore the corresponding weight $\boldsymbol{M}_{n,f}$ by setting it to zero. However, this framework raises stability issues affecting the inference of the dropout parameters due to large-variance gradients, thus limiting $p_{n,f}$ to values smaller than 0.5. To tackle this problem, we leverage on the extension of *Variational Dropout* proposed in [29]. In this setting, the variance parameter is encoded in a new independent variable $\boldsymbol{P}_{n,f} = \alpha_{n,f}\boldsymbol{M}_{n,f}^2$, while the posterior distribution is optimized with respect to $(\boldsymbol{M}, \boldsymbol{P})$. Therefore, in order to minimize the cost function for large variance $\boldsymbol{P}_{n,f} \to \infty$ ($\alpha_{n,f} \to \infty$ i.e $p_{n,f} \to 1$), the value of the weight's magnitude must be controlled by setting to zero the corresponding parameter $\boldsymbol{M}_{n,f}$. As a result, by dropping out weights in the code, we sparsify the estimated spatial maps, thus better isolating relevant spatial sub-structures. Spatial correlations in the images are obtained thanks to the convolution operation detailed in Section 3.3.2.

3.5 Variational inference

We detailed in the previous sections the choices of priors and constraints that we apply to the spatio-temporal processes in order to plausibly model the data. To illustrate the overall formulation of the method, we provide in Figure 2 the graphical model over the M modalities in the case of imaging data. Naturally, this graph simplifies when we deal with scalar data as we don't need to account for any spatial dependence. To infer the time-shift parameter $\boldsymbol{\delta}$, the sets of parameters θ_m, θ_c, and ψ_m, as well as $\boldsymbol{Z}$, σ and ν, we need to jointly optimize the data evidence according to priors and constraints:

$$\log(p(\boldsymbol{Y}, \boldsymbol{V}, \mathcal{C}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu, \gamma)) = \sum_m \log(p(\boldsymbol{Y}^m, \mathcal{C}^m|\boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) + \sum_c \log(p(\boldsymbol{V}_{:c}, \mathcal{C}^c|\boldsymbol{\delta}, \nu_c, \gamma_c)). \tag{8}$$

We tackle the optimization of Equation (8) via stochastic variational inference. Following [9] and [24] we introduce approximations, $q_2(\boldsymbol{\Omega}^m)$ and $q_3(\boldsymbol{W}^m)$ in addition to $q_1(\boldsymbol{B}^m)$ in order to derive a lower bound $\boldsymbol{\mathcal{L}}_m$ for each modality. We recall that the temporal trajectories $\boldsymbol{S}^m$ and $\boldsymbol{U}$ are treated similarly as described in Section 3.3.1. We also note that the choice of distributions q_1, q_2 and q_3 is

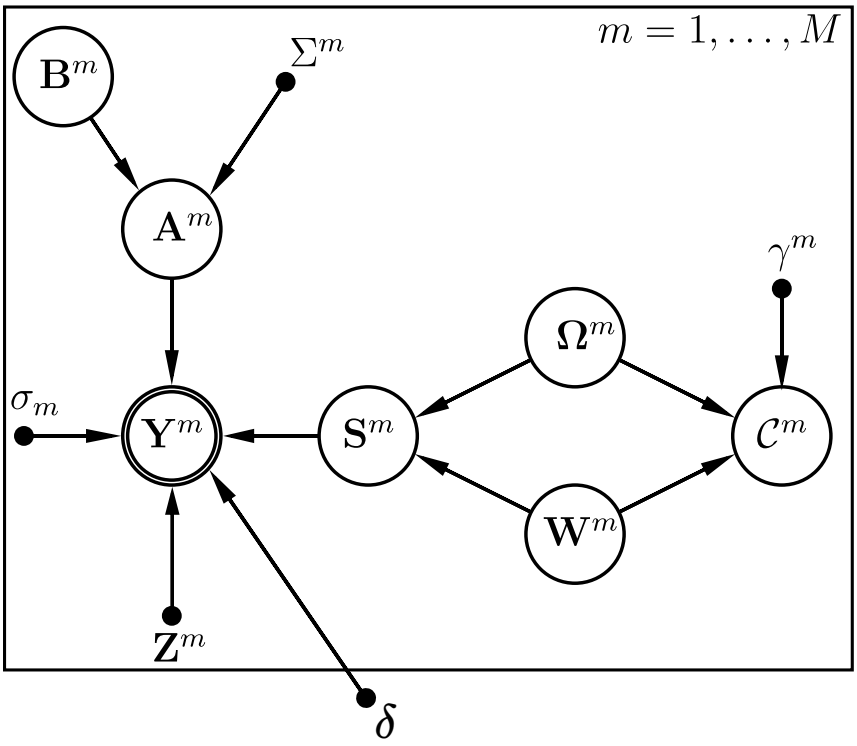


Figure 2: Graphical model for imaging data, $\boldsymbol{Y} = \{\boldsymbol{Y}^m\}$.

the same across modalities, while their parameters will be inferred independently. This leads to:

$$\begin{aligned}
\log(p(\boldsymbol{Y}^m, \mathcal{C}^m | \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) \geqslant & \mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}^m | \boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m))] \\
& + \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^m | \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m))] \\
& - \mathcal{D}[q_1(\boldsymbol{B}^m) || p(\boldsymbol{B}^m)] - \mathcal{D}[q_2(\boldsymbol{\Omega}^m) || p(\boldsymbol{\Omega}^m)] - \mathcal{D}[q_3(\boldsymbol{W}^m) || p(\boldsymbol{W}^m)], \\
\log(p(\boldsymbol{V}_{c:}, \mathcal{C}^c | \boldsymbol{\delta}, \nu_c, \gamma_c)) \geqslant & \mathbb{E}_{q_2,q_3}[\log(p(\boldsymbol{V}_{c:} | \boldsymbol{\Omega}^c, \boldsymbol{W}^c, \boldsymbol{\delta}, \sigma_c))] \\
& + \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^c | \boldsymbol{\Omega}^c, \boldsymbol{W}^c, \boldsymbol{\delta}, \gamma_c))] \\
& - \mathcal{D}[q_2(\boldsymbol{\Omega}^c) || p(\boldsymbol{\Omega}^c)] - \mathcal{D}[q_3(\boldsymbol{W}^c) || p(\boldsymbol{W}^c)]
\end{aligned} \tag{9}$$

Where $\mathcal{D}$ refers to the Kullback-Leibler (KL) divergence. Combining the lower bounds of the different modalities we obtain:

$$\log(p(\boldsymbol{Y}, \boldsymbol{V}, \mathcal{C} | \boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu, \gamma)) \geqslant \sum_m \boldsymbol{\mathcal{L}}_m + \sum_c \boldsymbol{\mathcal{L}}_c. \tag{10}$$

A detailed derivation of the lower bound is given in Appendix A.
The approximated distributions $q_2(\boldsymbol{\Omega}^m)$ and $q_3(\boldsymbol{W}^m)$ are factorized across GPs such that:

$$\begin{aligned} q_2(\boldsymbol{\Omega}^m) &= \prod_{n=1}^{N_m} q_2(\boldsymbol{\omega}^n)^m = \prod_{n=1}^{N_m}\prod_{j=1}^{N_{rf}} \mathcal{N}(\boldsymbol{R}_{n,j}, \boldsymbol{Q}_{n,j}^2)^m, \\ q_3(\boldsymbol{W}^m) &= \prod_{n=1}^{N_m} q_3(\boldsymbol{w}^n)^m = \prod_{n=1}^{N_m}\prod_{j=1}^{N_{rf}} \mathcal{N}(\boldsymbol{T}_{n,j}, \boldsymbol{V}_{n,j}^2)^m, \end{aligned} \tag{11}$$

where N_{rf} is the number of random features used for the projection in the spectral domain. Using Gaussian priors and approximations we introduced above, we can obtain a closed-form formula for the KL divergence. Moreover, the choice of prior and approximate posterior distribution for the maps of $\boldsymbol{B}^m$ leads to an approximation for the divergence $\mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)]$ detailed in [29]. This allows to analytically compute all the KL terms in our cost function. Formulas for the KL divergences are detailed in Appendix B.

Finally, we optimize the individual time-shifts $\boldsymbol{\delta} = \{\delta_p\}_{p=0}^P$, $\boldsymbol{Z}$, $\sigma = \{\sigma_m\}_{m=1}^M$, $\nu = \{\nu_c\}_{c=1}^C$ as well as the overall sets of spatio-temporal parameters $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^M \cup \{\theta_c\}_{c=1}^C$ and $\boldsymbol{\psi} = \{\psi_m\}_{m=1}^M$.

$$\begin{aligned} \boldsymbol{\theta} &= \{\boldsymbol{R}_{n:}^m, \boldsymbol{Q}_{n:}^m, \boldsymbol{T}_{n:}^m, \boldsymbol{V}_{n:}^m, l_n, n \in [1, N_m]\}_{m=1}^M \cup \{\boldsymbol{R}_{c:}, \boldsymbol{Q}_{c:}, \boldsymbol{T}_{c:}, \boldsymbol{V}_{c:}^c, l_c\}_{c=1}^C, \\ \boldsymbol{\psi} &= \{\boldsymbol{M}_{n:}^m, \boldsymbol{P}_{n:}^m, n \in [1, N_m]\}_{m=1}^M. \end{aligned} \tag{12}$$

Following [20] and using the reparameterization trick, we can efficiently sample from the approximated distributions q_1, q_2 and q_3 to compute the two expectation terms from (9) for each modality. We chose to alternate the optimization between the spatio-temporal parameters and the time-shift. We set γ_m to the minimum value that gives monotonic sources. This was done through multiple tests on data batches with different numbers of imaging features F_m and sources N_m. We empirically found that monotonicity was enforced when the magnitude of γ_m was in the order of $F_m \times N_m$. The threshold for the dropout probability above which we set a weight $\boldsymbol{B}_{n,f}^m$ to zero was fixed at 95% (i.e $\alpha = 19$), while the σ_m and ν_m were optimized during training along with the spatio-temporal parameters. The model is implemented and trained using the Pytorch library [30]. The complete experimental setting is detailed in Appendix C. We also provide a pseudo-code detailing the optimization procedure in Appendix D. In the following sections we will refer to our method as Monotonic Gaussian Process Analysis (MGPA).

4 Experiments and Results

In this section we first benchmark MGPA on synthetic data to demonstrate its reconstruction and separation properties while comparing it to standard sources separation methods. We finally apply our model on a large set of medical data from a publicly available clinical study, demonstrating the ability of our method to retrieve spatio-temporal processes relevant to AD, along with a time-scale describing the course of the disease.

4.1 Synthetic tests on spatio-temporal trajectory separation

For the synthetic tests we considered the case where the data is associated to a single imaging modality only. We tested MGPA on synthetic data generated as a linear combination of temporal functions and 3D activation maps at prescribed resolutions. The goal was to assess the method's ability to identify the spatio-temporal sources underlying the data. We benchmarked our method with respect to ICA, Non-Negative Matrix Factorization (NMF), and Principal Component Analysis (PCA), which were applied from the standard implementation provided in the Scikit-Learn library [31].

The benchmark was specified by defining a 10-folds validation setting, generating the data at each fold as a linear combination of temporal sources $\widetilde{\boldsymbol{S}}(\boldsymbol{t}) = [\widetilde{\boldsymbol{S}}_{:0}(\boldsymbol{t}), \widetilde{\boldsymbol{S}}_{:1}(\boldsymbol{t})]$, and spatial maps $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{A}}_{0:}, \widetilde{\boldsymbol{A}}_{1:}]$. The data was defined as $\boldsymbol{Y}_{p:} = \widetilde{\boldsymbol{S}}_{p:}(\boldsymbol{t}_p)\widetilde{\boldsymbol{A}} + \boldsymbol{\mathcal{E}}_{p:}$ over 50 time points $\boldsymbol{t}_p$, where $\boldsymbol{t}_p$ was uniformly distributed in the range $[0, 0.7]$, and $\boldsymbol{\mathcal{E}}_{p:} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The temporal sources were specified as sigmoid functions $\widetilde{\boldsymbol{S}}_{p,i}(\boldsymbol{t}_p) = 1/(1 + \exp(-\boldsymbol{t}_p + \alpha_i))$, while the spatial structures had dimensions $(30 \times 30 \times 30)$ such that $\widetilde{\boldsymbol{A}}_{i:} = \widetilde{\boldsymbol{B}}_{i:}\widetilde{\boldsymbol{\Sigma}}^i$. The $\widetilde{\boldsymbol{\Sigma}}^i$ were chosen as Gaussian convolution matrices with respective length-scale of $\lambda = 2$ mm and $\lambda = 1$ mm. The $\widetilde{\boldsymbol{B}}_{i:}$ were randomly sampled sparse 3D maps.

Variable selection. We applied our method by specifying an over-complete set of six sources with respective spatial length-scale of $\lambda = \{2, 2, 1, 1, 0.5, 0.5\ mm\}$. Figure 3 shows an example of the sparse maps obtained for a specific fold. The model prunes the signal for most of the maps, while retaining two sparse maps, $\boldsymbol{B_{0:}}$ and $\boldsymbol{B_{4:}}$, whose length-scale are $\lambda = 2$ mm and $\lambda = 1$ mm, thus correctly estimating the right number of sources and their spatial resolution. As it can be qualitatively observed in Figure 3, we notice that the estimated sparse code convolved with a Gaussian kernel matrix with $\lambda = 1$ mm is closer to its ground truth than the one convolved with a length-scale $\lambda = 2$ mm. According to our tests, sparse codes associated to high resolution details (low λ) are indeed more identifiable. On the contrary, the identifiability of images obtained via a convolution operator with larger kernels (large λ) is lower, since these maps can be equivalently obtained through the convolution of different sparse codes.

Sources separation. We observe in Table 1 that the lowest Mean-Squared Error (MSE) for the temporal sources reconstruction is obtained by MGPA, closely followed by ICA. Similarly, our model and ICA show the highest Structural Similarity (SSIM) score [38], which quantifies the image reconstruction accuracy with respect to the ground truth maps, while accounting for the inter-dependencies between neighbouring pixels. An example of image reconstruction from a sample fold is illustrated in Figure 4. In this standard benchmark, we note that MGPA leads to comparable results with respect to the state of the art. In the following section, we compare the models in the more challenging setting in which the time-line has to be estimated as well.

4.2 Synthetic tests on trajectory separation and time-reparameterization

In this test, we modify the experimental benchmark by introducing a further element of variability associated to the time-axis. The temporal and spatial sources were modelled following the same procedure as in Section 4.1, however the observations were mixed along the temporal axis. To do so we generated longitudinal data as $\boldsymbol{Y}_{p,j,:} = \widetilde{\boldsymbol{S}}_{p:}(\boldsymbol{t})\widetilde{\boldsymbol{A}} + \boldsymbol{\mathcal{E}}_{j:}$, by sampling between 1 and 10

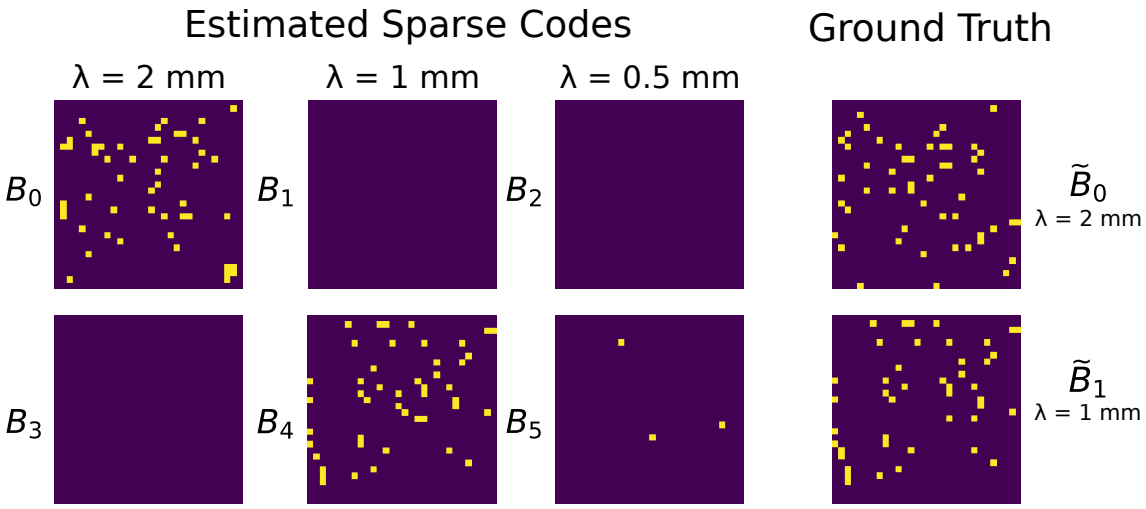


Figure 3: Slices extracted from the six sparse codes and the ground truth. Blue: Rejected points. Yellow: Retained points.

Table 1: MSE and SSIM between respectively the ground truth temporal and spatial sources with respect to the ones estimated by the different methods.

	Temporal (MSE)	Spatial (SSIM)
MGPA	$(\mathbf{8} \pm \mathbf{4}).\mathbf{10^{-5}}$	$\mathbf{98\%} \pm \mathbf{1}$
ICA	$(6 \pm 3).10^{-4}$	$97\% \pm 2$
NMF	$(3 \pm 2).10^{-2}$	$40\% \pm 17$
PCA	0.44 ± 10^{-3}	$15\% \pm 1$

images per time-point and randomly re-arranging them along the time-axis (cf. time-shift t_p of each observation at initialization in Figures 5 and 6, panel "Time-Shift"). The goal was to assess the sources separation performances of MGPA when the time-line is unknown. The experiment was run on 10 folds and Figures 5 and 6 illustrate the sources estimation for two different folds. We present these two figures to demonstrate how the time-shift inference affects the temporal sources reconstruction. Since the model is agnostic of a time-scale, we note that the time-shift may have a different range than the original time-axis. However, its relative ordering should be consistent with the original time points. We fitted a linear regression model over the 10 folds between the original time and the estimated time-shift parameter, and obtained an average R^2 coefficient of 0.98 with a standard deviation of 0.005 (cf. Table 2). This is illustrated for two different folds in the Time-Shift panel of Figures 5 and 6, where we observe a strong linear correlation with the original time-line, meaning that the algorithm correctly re-ordered the data with respect to the original time-axis. However, we notice in Table 2 that the MSE of the temporal sources significantly increased, due to the additional difficulty brought by the time-shift estimation. Indeed, in order to reconstruct the temporal signal we need to perfectly re-align hundreds of observations. This is

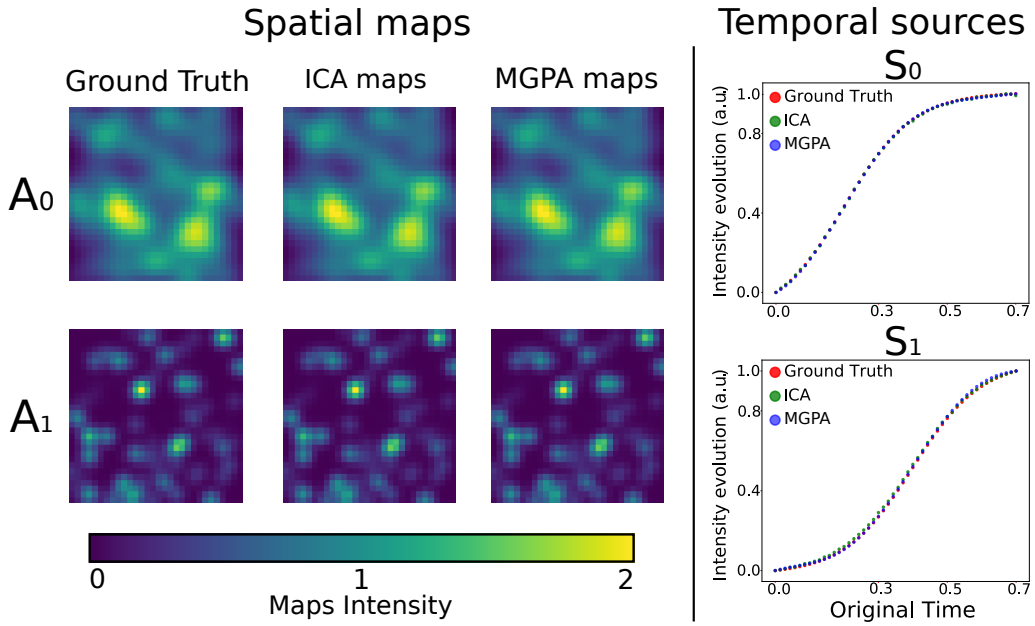


Figure 4: Spatio-temporal reconstruction when inference on the time-line is not required. Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm), the maps estimated by ICA, and the ones estimated by MGPA. Temporal sources: Ground truth temporal sources (red) along with sources estimated by ICA (green) and MGPA (blue).

Table 2: MSE and SSIM between respectively the ground truth temporal and spatial sources with respect to the ones estimated by MGPA. R^2 coefficient of the linear regression between the original time-line and the estimated time-shift.

	Temporal (MSE)	Spatial (SSIM)	R^2
MGPA	$\mathbf{(2 \pm 0.8).10^{-2}}$	$\mathbf{95\% \pm 4}$	$\mathbf{0.98 \pm 0.005}$

the case in Figure 5 (optimal reconstruction result), where the time-shift is highly correlated with the original time-line, allowing to distinguish every single observation and reconstruct the original temporal profiles. Whereas in Figure 6 (sub-optimal reconstruction result), the estimated time-shift doesn't exhibit a perfect fit, and generally underestimates the time-reparameterization for the later and earlier time points. This is related to the challenging setting of reconstructing the time-line identified by the original temporal sources. Indeed, we observe that $\boldsymbol{S}_{:0}$ reaches a plateau for early time points, while $\boldsymbol{S}_{:1}$ is flat for later ones. This behaviour increases the difficulty of differentiating time points with low signal differences. As a result, it impacts the time-shift optimization and adds variability to the time-shift estimation performances, thus deteriorating the reconstruction of the temporal sources over the 10 folds compared to the previous benchmark. The spatial sources estimation remains comparable to the one without time-shift both quantitatively, with an average

SSIM of 95%, and qualitatively, as shown in Figures 5 and 6. Within this setting, ICA, NMF and PCA poorly perform as they can't reconstruct the time-line. Results obtained using these three methods are provided in Appendix E.

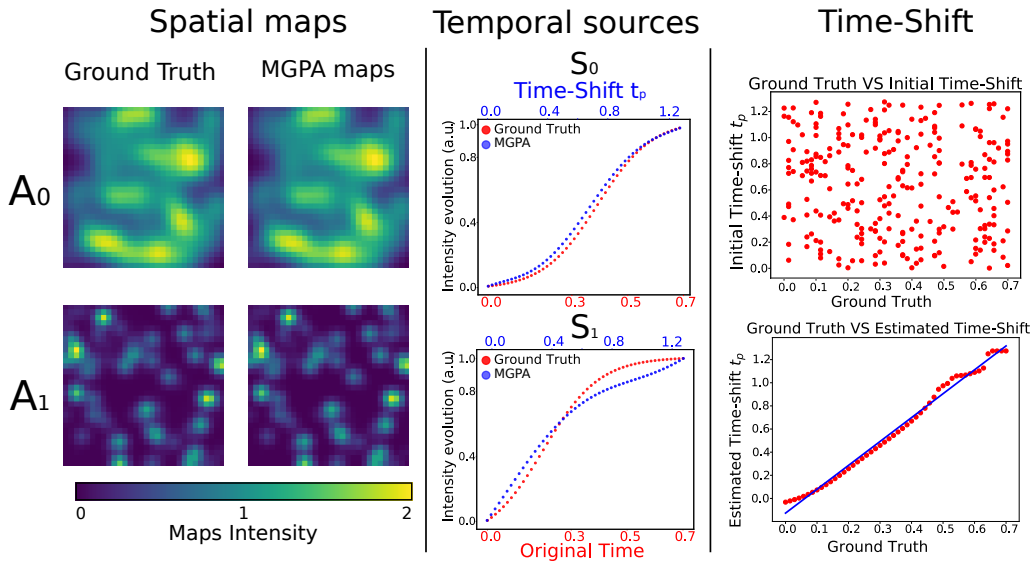


Figure 5: Spatio-temporal reconstruction when inference on the time-line is required. Optimal reconstruction result. Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm) and estimated spatial sources. Temporal sources: In red the original temporal sources, in blue the estimated temporal sources. Time-Shift: Time-shift t_p of each image at initialization (top), and after estimation (bottom). In blue, linear fit with the ground truth.

4.3 Application to spatio-temporal brain progression modeling

4.3.1 Data processing

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.

We selected a cohort of 544 amyloid positive subjects of the ADNI database composed of 103 controls (NL), 164 Mild Cognitive Impairment (MCI), 114 AD patients, 34 healthy individuals converted to MCI or to AD (NL converter) and 129 MCI converted to AD (MCI converter). The term amyloid positive refers to subjects whose amyloid level in the cerebrospinal fluid (CSF) is below the nominal cutoff of 192 pg/ml. Conversion to MCI or AD was determined using the last follow-up available information. We provide in Table 3 socio-demographic and clinical information across the different groups.

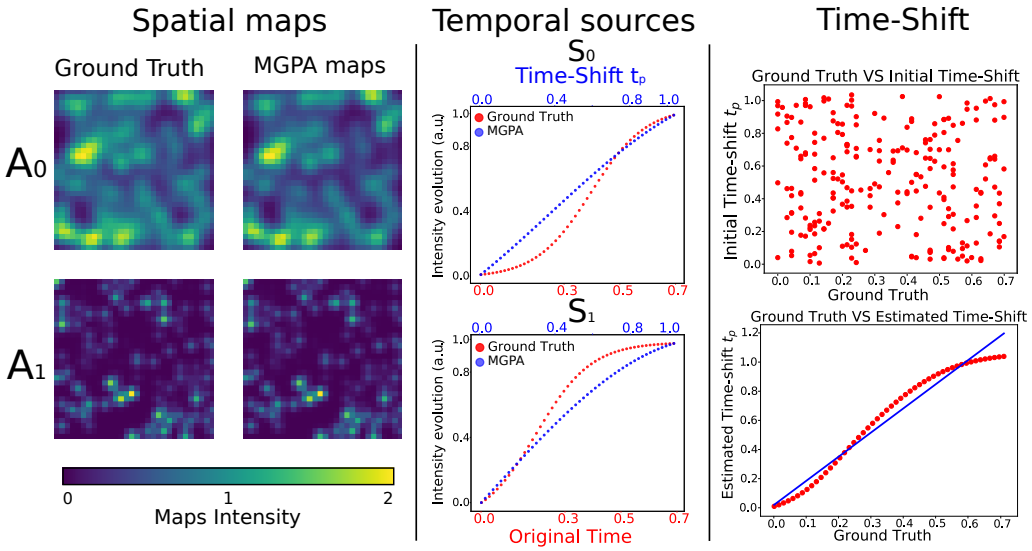


Figure 6: Spatio-temporal reconstruction when inference on the time-line is required. Sub-optimal reconstruction result. Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm) and estimated spatial sources. Temporal sources: In red the original temporal sources, in blue the estimated temporal sources. Time-Shift: Time-shift t_p of each image at initialization (top), and after estimation (bottom). In blue, linear fit with the ground truth.

Table 3: Baseline socio-demographic and clinical information for study cohort. Average values and standard deviation in parenthesis. NL: normal individuals, NL converter: normal subjects who converted to MCI or to AD, MCI: mild cognitive impairment, MCI converter: MCI subjects who converted to AD, AD: Alzheimer's patients. ADAS13: Alzheimer's Disease Assessment Scale-cognitive subscale, 13 items. FAQ: Functional Assessment Questionnaire. FDG: (18)F-fluorodeoxyglucose Positron Emission Tomography (PET) imaging. AV45: (18)F-florbetapir Amyloid PET imaging.

Group	NL	NL converter	MCI	MCI converter	AD
N	103	34	164	129	114
Age	73 (6)	78 (5)	73 (7)	73 (7)	74 (8)
Education (yrs)	16.3 (3)	16 (3)	15.7 (3)	16 (3)	15.6 (3)
ADAS13	9.1 (4.4)	11.4 (4.3)	14.6 (5.5)	20.4 (6.5)	31.6 (8.5)
FAQ	0.3 (0.7)	0.2 (0.6)	1.9 (2.8)	5.0 (4.6)	13.5 (6.9)
Entorhinal (cm^3)	3.8 (0.5)	3.5 (0.5)	3.6 (0.6)	3.2 (0.7)	2.8 (0.6)
Hippocampus (cm^3)	7.4 (0.9)	6.9 (0.7)	6.9 (0.9)	6.4 (0.9)	5.9 (0.8)
Ventricles (cm^3)	31 (16)	42 (21)	39 (23)	40 (19)	48 (23)
Whole brain (cm^3)	1033 (104)	1019 (91)	1058 (103)	1037 (102)	1005 (115)
FDG	1.3 (0.1)	1.3 (0.1)	1.2 (0.1)	1.1 (0.1)	1.0 (0.1)
AV45	1.3 (0.2)	1.3 (0.1)	1.3 (0.2)	1.4 (0.2)	1.5 (0.2)

MRI, FDG-PET and AV45-PET of each individual were processed in order to obtain respectively, volumes of gray matter density, glucose uptake, and amyloid load in a standard anatomical space.

MRI processing protocol. Baseline MRI images were analyzed according to the SPM12 processing pipeline [1]. Each image was initially segmented into grey, white matter and CSF probabilistic maps. Grey matter images were used for the following analysis, normalized to a group-wise reference space via DARTEL [2], and modulated using the Jacobian determinant of the subject-to-template transformation. The subsequent modeling was carried out on the normalised images at the original spatial resolution.

PET processing protocol. Individuals' baseline PET images were initially affinely aligned to the corresponding MRI. After scaling the intensities to the cerebellum, the images were normalized to the grey matter template obtained with DARTEL and smoothed with a FWHM parameter of 4.55.

Images have dimension $102 \times 130 \times 107$ before vectorization, leading to $1,418,820$ spatial features per patient. These spatial features represent for each voxel their gray matter concentration in the case of MRI images, their glucose metabolism for FDG-PET images, or their amyloid concentration for AV45-PET images. To exploit the ability of our model to automatically adapt to different spatial scales, we chose to keep the MRI images at their native resolution for the analysis, and thus do not perform additional smoohting to equalize to the PET FWHM. In addition to the imaging data of each patient, we also integrate the ADAS13 score assessed by clinicians. High values of this score indicate a decline of cognitive abilities. We consider three matrices $\boldsymbol{Y}^{MRI}$, $\boldsymbol{Y}^{FDG}$, and $\boldsymbol{Y}^{AV45}$ of dimension $(543 \times 1,418,820)$ containing the images of all the subjects, and a matrix $\boldsymbol{V}$ of dimension (543×1) containing their ADAS13 score. From now on we will refer to the data as the block diagonal matrix containing the four matrices $\boldsymbol{Y}^{MRI}, \boldsymbol{Y}^{FDG}$, $\boldsymbol{Y}^{AV45}$, and $\boldsymbol{V}$ as described in Section 3.2. We note that the analysis is performed by only considering a single scan per imaging modality and ADAS13 score for each patient. Therefore, the temporal evolution has to be inferred solely through the analysis of relative differences between the brain morphologies, glucose metabolisms, amyloid concentrations and cognitive abilities across individuals.

4.3.2 Model specification

We aim at showing how MGPA applied on the data extracted from the ADNI cohort is able to temporally re-align patients in order to describe AD progression in a plausible way, while detecting relevant spatio-temporal processes at stake in AD. The model estimates AD progression by relying on MR, FDG-PET, AV45-PET scans and ADAS13 score of each patient. The temporal sources $\boldsymbol{S}^{MRI}$ and $\boldsymbol{S}^{FDG}$ associated respectively to the loss of gray matter, and to the decrease of glucose uptake, are enforced to be monotonically decreasing. On the contrary, the temporal sources $\boldsymbol{S}^{AV45}$ and $\boldsymbol{U}_{:ADAS13}$, modeling respectively the evolution of amyloid concentration, and ADAS13 score, are enforced to be monotonically increasing. Since we don't consider any information about the disease stage of each individual before applying our method, all the observations are initialized at the same time reference $\tau = 0$. Therefore, as for the tests in Section 4.2, the time-shift reparameterization describes a relative re-ordering of the subjects not related to a specific time-unit. To decompose the imaging data we apply our model by specifying an over-complete basis of six sources with $\lambda = \{8, 8, 4, 4, 2, 2\ mm\}$, to cover both different scales and the associated variety of temporal

evolution. Due to the high-dimension of the data matrix, the computations were parallelized over six GPUs, and the model required eighteen hours to complete the training. Details on the model convergence during training are provided in Appendix F.

4.3.3 Estimated spatio-temporal brain dynamics

In Figure 7 we show the spatio-temporal processes retained by the model for each imaging modality. Interestingly, the model adapts to the spatial resolution of MRI and PET images. Indeed, we notice that the model accounts for the high-resolution of MRI images by retaining a source associated to the lowest length-scale ($\lambda = 2$ mm). Concerning PET data, we observe that the induced sparsity discards the highest resolution codes ($\lambda = 2$ mm) for both FDG and AV45, highlighting the ability of the model to adapt to the coarser resolution of the PET signal.

In the case of MRI data, two sources were retained at two different resolutions ($\lambda = 4$ mm and $\lambda = 2$ mm). Source $\boldsymbol{S}_4^{MRI}$ describes gray matter loss encompassing a large extent of the brain with a focus on cortical areas (see $\boldsymbol{A}_4^{MRI}$). We note that this map also targets subcortical areas such as the hippocampi, which are key regions of AD. Source $\boldsymbol{S}_2^{MRI}$ ($\lambda = 4$ mm) indicates a mild decrease of gray matter which accelerates in the latest stages of the disease, and targets the temporal poles (see $\boldsymbol{A}_2^{MRI}$). It is interesting to notice that this differential pattern of gray matter loss also affects the parahippocampal region, whose atrophy is known to be prominent in AD [11]. These results underline the complex evolution of brain atrophy, and the ability of the model to disentangle spatio-temporal processes mapping different regions involved in the pathology [3, 13]. Concerning the spatio-temporal processes extracted from the FDG-PET data, we see on Figure 7 that the model retained two sources at the coarsest resolutions ($\lambda = 8$ mm). Source $\boldsymbol{S}_1^{FDG}$ indicates a pattern of hypometabolism that tends to plateau and which involves most of the brain regions, thus describing a global effect of the pathology on the glucose uptake. Source $\boldsymbol{S}_0^{FDG}$ describes a linear pattern of hypometabolism targeting areas such as the precuneus and the parietal lobe, which are known to be strongly affected during the evolution of the disease [5]. Finally, the model extracted two spatio-temporal sources from the AV45-PET data at two different resolutions ($\lambda = 8$ mm and $\lambda = 4$ mm). We observe that source $\boldsymbol{S}_2^{AV45}$ highlights an increase of amyloid deposition mapping a large extent of the brain, such as the parietal and frontal lobes as well as temporal areas, thus concurring with clinical evidence [34]. Similarly to the FDG-PET processes, we have a source $\boldsymbol{S}_0^{AV45}$ exhibiting a differential pattern of amyloid deposition targeting mostly frontal, temporal, occipital areas and precuneus.

The estimated spatio-temporal processes can be combined to obtain an estimated evolution $\boldsymbol{S}^m\boldsymbol{A}^m$ of the brain along the time-shift axis for each modality. In Figure 8, we show the ratio $|\boldsymbol{S}^m_{p:}\boldsymbol{A}^m - \boldsymbol{S}^m_{0:}\boldsymbol{A}^m|/\boldsymbol{S}^m_{0:}\boldsymbol{A}^m$ between the image predicted at four time-points t_p and the image predicted at t_0 for the three imaging modalities. This allows us to visualize the trajectory of a brain going from a healthy to a pathological state in terms of atrophy, glucose metabolism and amyloid load according to our model.

Finally, we also applied ICA, NMF and PCA on the ADNI data, showing that the associated results are characterized by poor interpretability and high variability. The complete experimental setting and results are detailed in Appendix G.

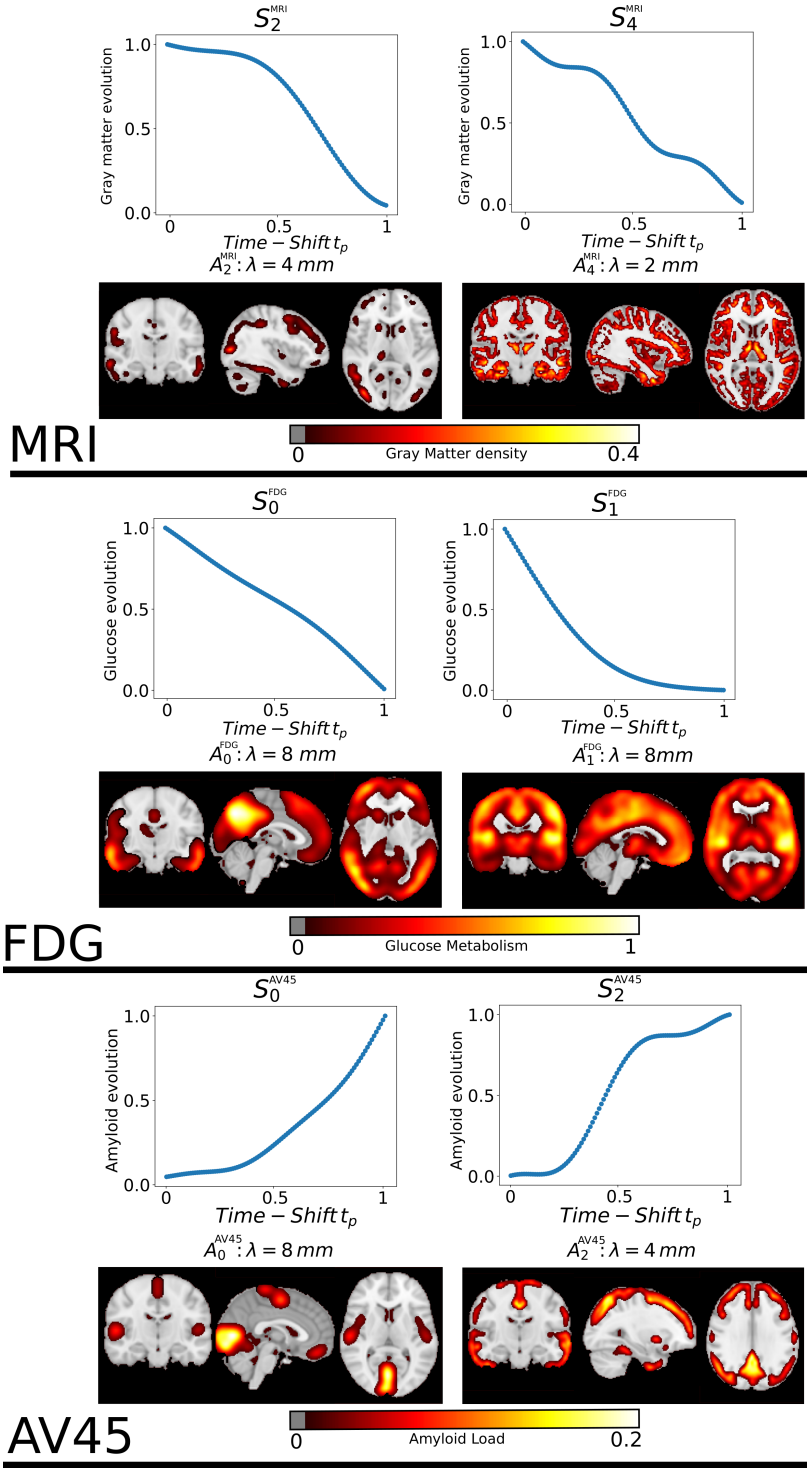


Figure 7: Estimated spatio-temporal processes for the three imaging modalities. The time-scale was re-scaled to the arbitrary range [0, 1].

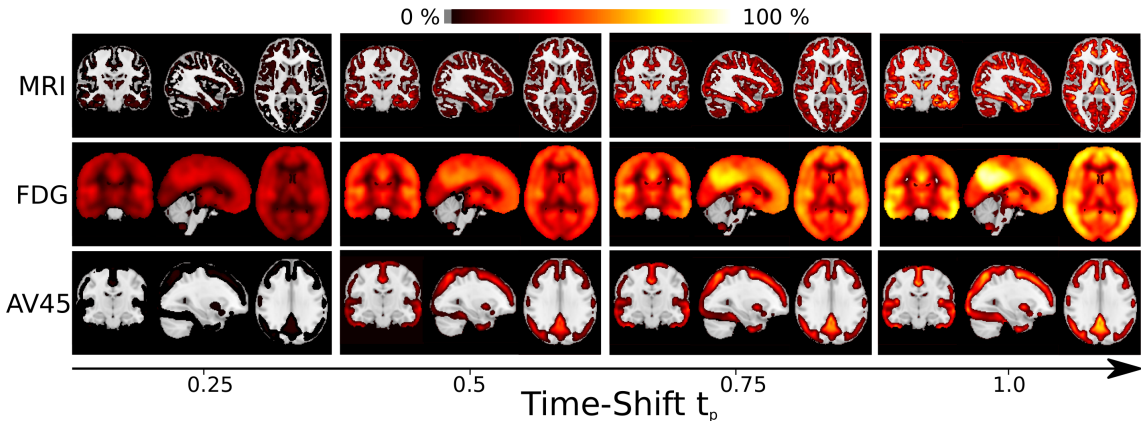


Figure 8: Ratio between the model prediction at time t_p and the prediction at t_0 for the three imaging modalities. The time-scale was re-scaled to the arbitrary range [0, 1].

4.3.4 Model Consistency

To verify the plausibility of the fitted model, we compare in Figure 9 the concentration predicted by the model and the raw concentration measures in different brain areas for the three imaging modalities. We observe a decrease of gray matter and glucose metabolism as we progress along the estimated time-line, allowing to relate large time-shift values to lower gray matter density and glucose uptake. Moreover, we notice the agreement between the predictions made by the model (in blue) and the raw concentration measures (in red). In the case of AV45 data there is only a mild increase of amyloid load according to the model, probably due to the fact that the subjects selected in the cohort are already amyloid positive. As a result, they already show a high baseline amyloid level concentration, close to plateau levels.

In Figure 10, we show the estimated GP $\boldsymbol{U}_{:ADAS13}$. We observe that the model is able to plausibly describe the evolution of this cognitive score, while demonstrating a larger variability than in the case of imaging modalities.

4.3.5 Plausibility with respect to clinical evidence

We assessed the clinical relevance of the estimated time-shift by relating it to independent medical information which were not included in the model during training. To this end, we compared the estimated time-shift to ADAS11, MMSE and FAQ scores. High values of ADAS11 and FAQ or low values of MMSE indicate a decline of performances. We show in Figure 11 that the estimated time-shift correlates with a decrease of cognitive and functional abilities. In particular, a cubic model slightly better describes the relationship between ADAS11 and the time-shift (according to BIC and AIC), with a significance for the cubic coefficient of $p = 0.04$. Concerning MMSE and FAQ, quadratic and linear models were almost equivalent; the significance of the linear coefficients was $p < 0.01$, while the quadratic coefficient was never significant. Pearson correlation coefficients for ADAS11, FAQ and MMSE were respectively of 0.49, 0.41, and -0.45, with corresponding p-values $p < 0.01$.

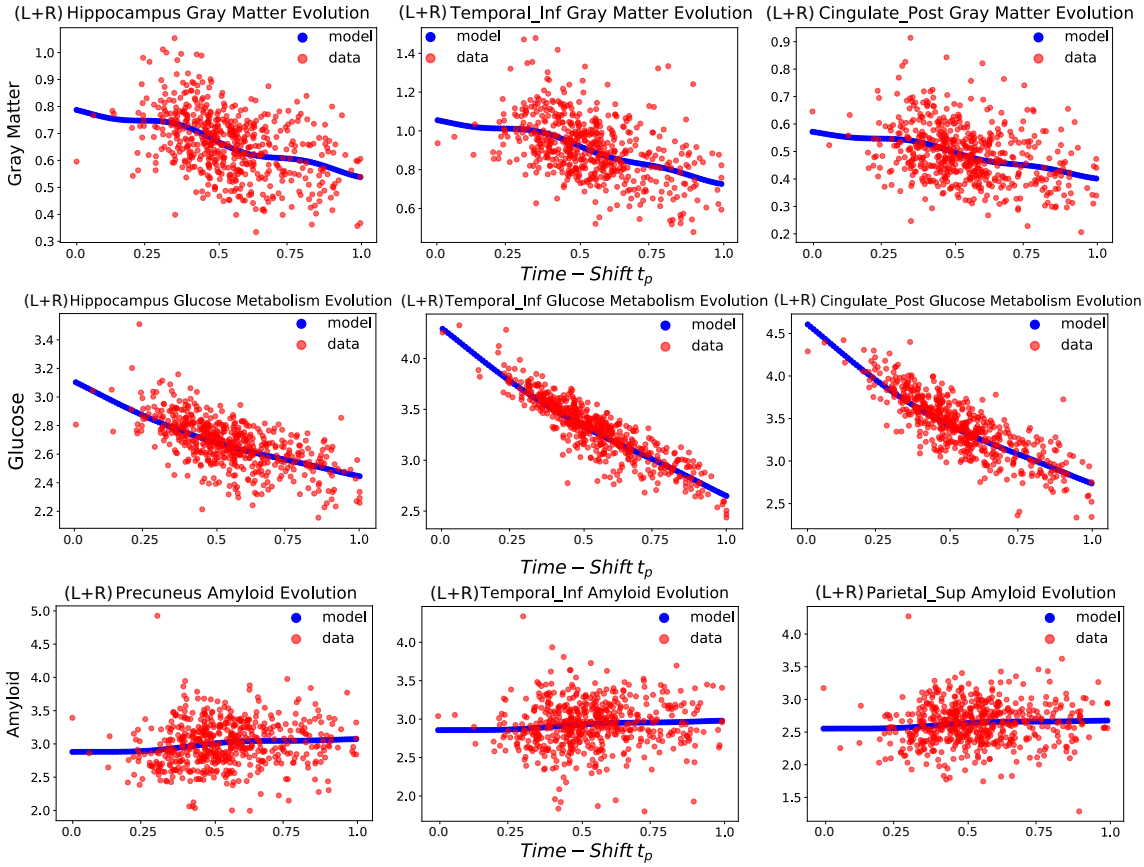


Figure 9: Model prediction averaged on specific brain areas (blue line), and observed values (red dots), along the estimated time-line for the three imaging modalities. L and R respectively stand for left and right. The time-scale was re-scaled to the arbitrary range [0, 1].

The box-plot of Figure 12 shows the time-shift distribution across clinical groups. We observe an increase of the estimated time-shift when going from healthy to pathological stages. The high uncertainty associated to the MCI group is due to the broad definition of this clinical category, which includes subjects not necessarily affected by dementia. We note that MCI subjects subsequently converted to AD (MCI converter) exhibit higher time-shift than the clinically stable MCI group, highlighting the ability of the model to differentiate between conversion status. A similar distinction can be noticed between NL and NL converter groups. We found significant differences between median time-shift for NL-NL converter, MCI-MCI converter and MCI converter-AD (comparisons $p < 0.01$, Figure 12). It is also important to recall that this result is obtained from the analysis of a single scan per imaging modality and ADAS13 score for each patient.

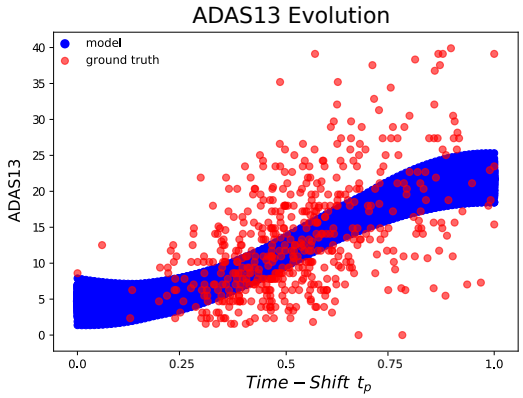


Figure 10: Model prediction of the ADAS13 score (blue line), and observed values (red dots) along the estimated time-line. The time-scale was re-scaled to the arbitrary range [0, 1].

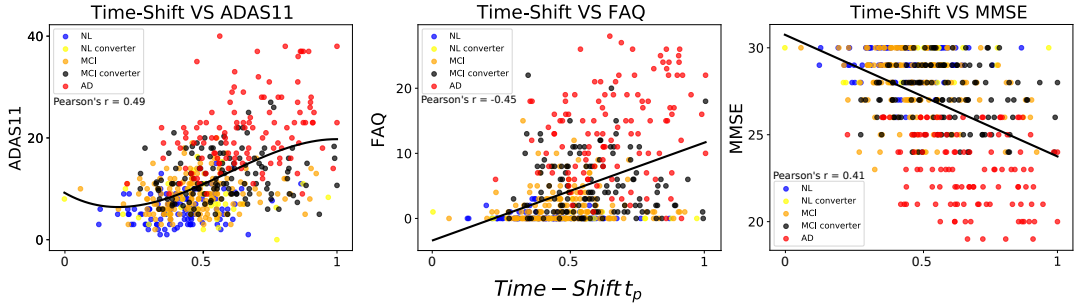


Figure 11: Evolution of the ADAS11 (left), FAQ (middle) and MMSE (right) along the estimated time-line. The time-scale was re-scaled to the arbitrary range [0, 1].

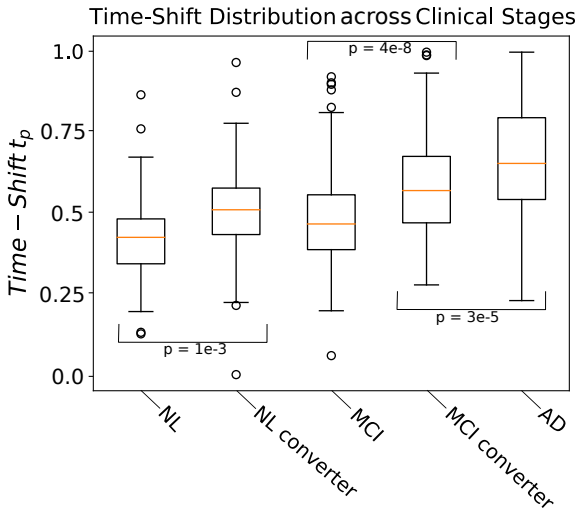


Figure 12: Distribution of the time-shift values over the different clinical stages. The time-scale was re-scaled to the arbitrary range [0, 1].

5 Discussion

We presented a generative approach to spatio-temporal disease progression modeling based on matrix factorization across temporal and spatial sources. The proposed application on a large set of medical images shows the ability of the model to disentangle relevant spatio-temporal processes at stake in AD, along with an estimated time-scale related to the disease evolution.

The model was compared to standard methods such as ICA, NMF and PCA since they perform blind source separation similarly to our method. This allowed us to demonstrate the advantages of building more complex approaches such as MGPA for the problem we tackle in this work. Concerning the comparison with the state of the art in disease progression modelling, to the best of our knowledge the two closest approaches are [27] and [21]. However, these two methods are specifically designed for modelling data defined on brain surfaces. On the contrary, our method aims at progression modeling using full 3D volumetric information. The data dimension we tackle is thus an order of magnitude greater than the one of [27] and [21], preventing these methods to scale to the spatial geometry of our data.

There are several avenues of improvement for the proposed approach. We found that the optimization is highly sensitive to the initialization of the spatial sources. This is typical of such complex non-convex problems, and requires further investigations to better control the algorithm convergence. More generally, the problem of source separation tackled in this work is intrinsically ill-posed, as the given data can be explained by several solutions. This was illustrated for example in our tests on synthetic data (Section 4.2), where the identification of the sources was more challenging in the case of coarse resolution codes and of flat temporal sources. We note however that this issue is general, and intrinsic to the problem of disease progression modeling.

Indeed, identifiability ultimately remains a critical issue when training the model. Concerning the spatio-temporal parameters, their number is extremely high due to the fact that we scale our method to 3D volumetric images. Estimating a single spatial source from a single modality requires to estimate the mean and variance of its sparse code, i.e $1,418,820 \times 2 = 2,837,640$ parameters. In practice, hypotheses are explicitly introduced to reduce the number of effective parameters. For instance, the convolution of the spatial maps using Gaussian kernels allows to enforce smoothness, and thus reduces the number of effective degrees of freedom via spatial correlation across the related parameters. This is equivalent to the regularization applied to image registration problems, in which the number of parameters is of the same order of magnitude than in our setting. Moreover, our sparsity constraint allows to sensibly reduce the number of parameters at test time. Indeed, after training, the sparse codes of the MRI sources have $2,213,359$ non-zero elements instead of $17,025,840$, which amounts in 87% reduction in the number of parameters. In the case of the FDG-PET and AV45-PET sparse codes, the number of non-zero elements at test time is respectively of $9,023,695$ and $1,362,067$, which is equivalent to a reduction in the number of parameters of 53% and 92%. Nonetheless, this high number of parameters still remains a factor of potential convergence issues during the parameters estimation procedure. We present graphs in Appendix F showing the evolution of the different terms composing the cost function during training. These figures show convergence profiles typical of those obtained with stochastic variational inference schemes, such as with Variational Autoencoders or Bayesian Neural Networks. Moreover, the stability of the solution has been ensured through multiple runs of the model. Finally, as mentioned in Section 3.4, the *Variational Dropout* framework leads to stability issues affecting inference, which are mostly due to the use of an improper prior. This problem may motivate the identification of alternative ways to induce sparsity on the spatial maps.

In this work, we modeled the time-shift of each subject as a translation with respect to a common temporal reference. However, since pathological trajectories are different across individuals, it would be valuable to account for individual speed of progressions by introducing a scaling effect, as it has been proposed for example in [21, 36]. This was not in the scope of the current study, as we focused on the analysis of cross-sectional data, thus having only one data point per subject. Therefore, one of the main extensions of this model will be the integration of longitudinal data for each individual, which will allow a more specific time-reparameterization.

Our noise model for the reconstruction problem of Equation 2 is homoscedastic and i.i.d. Gaussian with zero mean. For this reason, data variability for the entire image is encoded by the variance parameter of the Gaussian noise. Similarly as in standard regression problems, this modelling choice has been motivated to promote simplicity of the model and computational efficiency. However, around 40% of the values in the brain images do not provide relevant information as they represent zero and constant background areas. For this reason, during training, the model can perfectly fit this background and increases its confidence on the overall regression solution, thus lowering the value of the noise variance σ_m (cf Figure 9). This is in contrast to what we observe with the ADAS13 data (cf Figure 10), where the problem corresponds to standard univariate regression. A potential way to fix this issue could be to train the model only on non-zero image areas, or by implementing an heteroscedastic noise model. However, this latter solution may further increase the number of model parameters.

The modeling results are also sensitive to the specification of the spatio-temporal processes priors. In our case, the monotonicity constraint imposed to the GPs may be too restrictive to completely capture the complexity of the progression of neurodegeneration. From a clinical point of view, the model could also benefit from the integration of data measuring the concentration of Tau protein via PET imaging, in order to quantify key neurobiological processes associated to AD [17].

In order to guarantee that all the subjects belong to the same pathological trajectory due to AD, the model has only been applied to a cohort of amyloid positive subjects. However, this choice restricts the dynamics of evolution that we could estimate. Indeed, only considering these subjects narrows down the time-line of the pathology, as we study patients at potentially advanced disease stages. Therefore, it would be interesting in a future work to apply the model on a cohort including amyloid negative subjects, to model the brain dynamics over the whole disease natural history. This extension would require to define a proper methodology for disentangling sub-trajectories associated, for example with normal ageing and different pathological subtypes [22, 37, 40]. Moreover, we know that many patients diagnosed with AD can be associated to mixed pathologies such as vascular disease or Lewy bodies. Therefore, a potential clinical application of our method could be to investigate if the spatio-temporal dynamics estimated by MGPA are able to disentangle the contribution of each comorbidity.

Assessment of clinical plausibility of MGPA on the ADNI must be corroborated by further validation on independent datasets. Therefore, in a future work, we wish to validate the model on different cohorts to demonstrate its generalization properties. The validation step for each subject would be done by estimating the time-point minimizing the cost between the images of each tested individual, and the image progression model previously estimated on ADNI. The estimated time-shift would provide a measure of the pathological stage of the individual with respect to the modelled trajectory, and could be then compared with the clinical diagnosis of the subject, allowing to test the reliability of our model. This additional validation step could ultimately allow to use the model as a diagnostic instrument of AD. This validation would require an important effort in terms of data harmonisation across multiple cohorts, as well as in terms of clinical interpretation. For this reason, this work will be part of a subsequent publication.

We planned to release the source-code along with instructions in order for the model to be used by a large audience. It will be available as a complementary tool on the platform `http://gpprogressionmodel.inria.fr/`, which already offers a simple front-end to Gaussian Process Progression model.

6 Acknowledgements

This work has been supported by the French government, through the UCA^{JEDI} Investments in the Future project managed by the National Research Agency (ref.n ANR-15-IDEX-01), the grant AAP Sant 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) and DOD ADNI. ADNI is funded by the National Institute on Aging, the National

Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company;CereSpir, Inc.;Cogstate;Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.;Lumosity;Lundbeck;Merck & Co., Inc.; Meso Scale Diagnostics, LLC.;NeuroRx Research; Neurotrack Technologies;Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging;Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] J. Ashburner and K. J. Friston. Voxel-based morphometry–the methods. *NeuroImage*, 11(6 Pt 1):805–821, Jun 2000.

[2] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95 – 113, 2007.

[3] Randall J. Bateman, Chengjie Xiong, Tammie L.S. Benzinger, Anne M. Fagan, Alison Goate, Nick C. Fox, Daniel S. Marcus, Nigel J. Cairns, Xianyun Xie, Tyler M. Blazey, David M. Holtzman, Anna Santacruz, Virginia Buckles, Angela Oliver, Krista Moulder, Paul S. Aisen, Bernardino Ghetti, William E. Klunk, Eric McDade, Ralph N. Martins, Colin L. Masters, Richard Mayeux, John M. Ringman, Martin N. Rossor, Peter R. Schofield, Reisa A. Sperling, Stephen Salloway, and John C. Morris. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *New England Journal of Medicine*, 367(9):795–804, 2012. PMID: 22784036.

[4] M. Bilgel, B. Jedynak, D. F. Wong, S. M. Resnick, and J. L. Prince. Temporal Trajectory and Progression Score Estimation from Voxelwise Longitudinal Imaging Measures: Application to Amyloid Imaging. *Inf Process Med Imaging*, 24:424–436, 2015.

[5] R. K. Brown, N. I. Bohnen, K. K. Wong, S. Minoshima, and K. A. Frey. Brain PET in suspected dementia: patterns of altered FDG metabolism. *Radiographics*, 34(3):684–701, 2014.

[6] E. Bullmore, J. Fadili, V. Maxim, L. Sendur, B. Whitcher, J. Suckling, M. Brammer, and M. Breakspear. Wavelets and functional magnetic resonance imaging of the human brain. *NeuroImage*, 23 Suppl 1:S234–249, 2004.

[7] V. D. Calhoun, J. Liu, and T. Adali. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, 45(1 Suppl):S163–172, Mar 2009.

[8] Pierre Comon. Independent Component Analysis, a new concept? *Signal Processing*, 36:287–314, April 1994.

[9] Kurt Cutajar, Edwin V. Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep Gaussian processes. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[10] Michael C. Donohue, Hlne Jacqmin-Gadda, Mlanie Le Goff, Ronald G. Thomas, Rema Raman, Anthony C. Gamst, Laurel A. Beckett, Clifford R. Jack, Michael W. Weiner, Jean-Franois Dartigues, and Paul S. Aisen. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10(5, Supplement):S400 – S410, 2014.

[11] C. Echavarri, P. Aalten, H. B. Uylings, H. I. Jacobs, P. J. Visser, E. H. Gronenschild, F. R. Verhey, and S. Burgmans. Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Struct Funct*, 215(3-4):265–271, Jan 2011.

[12] H. M. Fonteijn, M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, and D. C. Alexander. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3):1880–1889, Apr 2012.

[13] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol*, 6(2):67–77, Feb 2010.

[14] K. Hackmack, F. Paul, M. Weygandt, C. Allefeld, and J. D. Haynes. Multi-scale classification of disease using structural MRI and wavelet transform. *NeuroImage*, 62(1):48–58, Aug 2012.

[15] C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*, 9(1):119–128, Jan 2010.

[16] B. M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, and J. L. Prince. A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort. *NeuroImage*, 63(3):1478–1486, Nov 2012.

[17] F. Kametani and M. Hasegawa. Reconsideration of Amyloid Hypothesis and Tau Hypothesis in Alzheimer's Disease. *Front Neurosci*, 12:25, 2018.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[19] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *CoRR*, abs/1506.02557, 2015.

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[21] Igor Koval, Jean-Baptiste Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stéphanie Allassonnière, and Stanley Durrleman. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In *Medical Image Computing and Computer Assisted Intervention*, Medical Image Computing and Computer Assisted Intervention, Quebec City, Canada, September 2017.

[22] M. Lorenzi, X. Pennec, G. B. Frisoni, and N. Ayache. Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. *Neurobiol. Aging*, 36 Suppl 1:42–52, Jan 2015.

[23] M. Lorenzi, G. Ziegler, D. C. Alexander, and S. Ourselin. Efficient Gaussian Process-Based Modelling and Prediction of Image Time Series. *Inf Process Med Imaging*, 24:626–637, 2015.

[24] Marco Lorenzi and Maurizio Filippone. Constraining the dynamics of deep probabilistic models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3233–3242, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[25] Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in alzheimer's disease. *NeuroImage*, 2017.

[26] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, July 1989.

[27] R. V. Marinescu, A. Eshaghi, M. Lorenzi, A. L. Young, N. P. Oxtoby, S. Garbarino, S. J. Crutch, and D. C. Alexander. DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage*, 192:166–177, 05 2019.

[28] A. F. Marquand, M. Brammer, S. C. Williams, and O. M. Doyle. Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage*, 92:298–311, May 2014.

[29] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2498–2507, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

[33] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 645–652, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[34] K. M. Rodrigue, K. M. Kennedy, and D. C. Park. Beta-amyloid deposition and the aging brain. *Neuropsychol Rev*, 19(4):436–450, Dec 2009.

[35] Yunus Saatçi. Scalable inference for structured gaussian process models, 2011.

[36] Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *NIPS*, pages 2404–2412, 2015.

[37] Raphaël Sivera, Hervé Delingette, Marco Lorenzi, Xavier Pennec, and Nicholas Ayache. A model of brain morphological changes related to aging and Alzheimer's disease from cross-sectional assessments. *NeuroImage*, 2019.

[38] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 13(4):600–612, 2004.

[39] J. L. Whitwell. Progression of atrophy in Alzheimer's disease and related disorders. *Neurotox Res*, 18(3-4):339–346, Nov 2010.

[40] A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, J. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonca, S. Sorbi, J. D. Warren, S. Crutch, N. C. Fox, S. Ourselin, J. M. Schott, J. D. Rohrer, and D. C. Alexander. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun*, 9(1):4273, 10 2018.

[41] A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, and D. C. Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(Pt 9):2564–2577, Sep 2014.

[42] A. L. Young, N. P. Oxtoby, J. Huang, R. V. Marinescu, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, and D. C. Alexander. Multiple Orderings of Events in Disease Progression. *Inf Process Med Imaging*, 24:711–722, 2015.

Appendix A.

In this Appendix, we detail the complete derivation of the lower bound.

$$\begin{aligned}\log(p(\boldsymbol{Y}^m, \mathcal{C}^m | \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m) p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{S}^m, \frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}|\boldsymbol{\delta}, \gamma) d\boldsymbol{B}^m d\boldsymbol{S}^m\Big] \\ &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m) p(\boldsymbol{B}^m) \\ &\qquad p(\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}|\boldsymbol{S}^m, \boldsymbol{\delta}, \gamma) p(\boldsymbol{S}^m) d\boldsymbol{B}^m d\boldsymbol{S}^m\Big].\end{aligned}$$

By observing that $\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}$ is completely identified by $\boldsymbol{S}^m$, the equation can be written as:

$$\begin{aligned}\log(p(\boldsymbol{Y}^m, \mathcal{C}^m | \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m) p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{S}^m) d\boldsymbol{B}^m d\boldsymbol{S}^m\Big].\end{aligned}$$

Similarly this derivation can be applied to $\log(p(\boldsymbol{V}_{:c}, \mathcal{C}^c | \boldsymbol{\delta}, \nu_c, \gamma_c))$.

$$\begin{aligned}\log(p(\boldsymbol{Y}^m, \mathcal{C}^m | \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m) p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{S}^m) d\boldsymbol{B}^m d\boldsymbol{S}^m\Big] \\ &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m) p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{\Omega}^m) p(\boldsymbol{W}^m) d\boldsymbol{B}^m d\boldsymbol{\Omega}^m d\boldsymbol{W}^m\Big] \\ &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m) p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{\Omega}^m) p(\boldsymbol{W}^m) \frac{q_1(\boldsymbol{B}^m) q_2(\boldsymbol{\Omega}^m) q_3(\boldsymbol{W}^m)}{q_1(\boldsymbol{B}^m) q_2(\boldsymbol{\Omega}^m) q_3(\boldsymbol{W}^m)} d\boldsymbol{B}^m d\boldsymbol{\Omega}^m d\boldsymbol{W}^m\Big] \\ &= \log\Big[\mathbb{E}_{q_1,q_2,q_3} \frac{p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m)}{q_1(\boldsymbol{B}^m) q_2(\boldsymbol{\Omega}^m) q_3(\boldsymbol{W}^m)} \\ &\qquad \frac{p(\boldsymbol{B}^m) p(\boldsymbol{\Omega}^m) p(\boldsymbol{W}^m)}{q_1(\boldsymbol{B}^m) q_2(\boldsymbol{\Omega}^m) q_3(\boldsymbol{W}^m)}\Big] \\ &\geq \mathbb{E}_{q_1,q_2,q_3}\Big(\log\Big[\frac{p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m) p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m)}{q_1(\boldsymbol{B}^m) q_2(\boldsymbol{\Omega}^m) q_3(\boldsymbol{W}^m)} \\ &\qquad \frac{p(\boldsymbol{B}^m) p(\boldsymbol{\Omega}^m) p(\boldsymbol{W}^m)}{q_1(\boldsymbol{B}^m) q_2(\boldsymbol{\Omega}^m) q_3(\boldsymbol{W}^m)}\Big]\Big) \\ &= \mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m))]\end{aligned}$$

$$+ \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^m|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m))]$$
$$- \mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)] - \mathcal{D}[q_2(\boldsymbol{\Omega}^m)||p(\boldsymbol{\Omega}^m)] - \mathcal{D}[q_3(\boldsymbol{W}^m)||p(\boldsymbol{W}^m)].$$

This derivation gives us the lower bound $\boldsymbol{\mathcal{L}}_m$ of a given modality m. The same technique can be used to derive a lower bound for $\log(p(\boldsymbol{V}_{c:}, \mathcal{C}^c|\boldsymbol{\delta}, \nu_c, \gamma_c))$, and by summation over m and c we obtain the lower bound of Equation 10 for $\log(p(\boldsymbol{Y}, \boldsymbol{V}, \mathcal{C}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu, \gamma))$.

Appendix B.

In this section we provide formulas for computing the three KL terms of the lower bound. The total KL divergences are:

$$\mathcal{D}[q_1(\boldsymbol{B})||p(\boldsymbol{B})] = \sum_m \mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)],$$
$$\mathcal{D}[q_2(\boldsymbol{\Omega})||p(\boldsymbol{\Omega})] = \sum_m \mathcal{D}[q_1(\boldsymbol{\Omega}^m)||p(\boldsymbol{\Omega}^m)] + \sum_c \mathcal{D}[q_1(\boldsymbol{\Omega}^c)||p(\boldsymbol{\Omega}^c)],$$
$$\mathcal{D}[q_3(\boldsymbol{W})||p(\boldsymbol{W})] = \sum_m \mathcal{D}[q_3(\boldsymbol{W}^m)||p(\boldsymbol{W}^m)] + \sum_c \mathcal{D}[q_3(\boldsymbol{W}^c)||p(\boldsymbol{W}^c)].$$

For ease of notation we will drop the m and c indices and will give formulas for a single modality. In [29], authors provide an approximation of the KL for the maps $\boldsymbol{B}$:

$$-\mathcal{D}[q_1(\boldsymbol{B})||p(\boldsymbol{B})] = \sum_{n,f} k_1 h(k_2 + k_3 \log(\alpha_{n,f})) - 0.5 \log(1 + \alpha_{n,f}^{-1}) - k_1,$$

where h is the sigmoid function and $k_1 = 0.63576,\ k_2 = 1.87320,\ k_3 = 1.48695$.

In the case of $\boldsymbol{\Omega}$ and $\boldsymbol{W}$, we've seen that they have Gaussian priors and approximations which are detailed in Sections 3.3.1 and 3.5. As a result we can obtain closed-form formulas for their KL, leading to:

$$\mathcal{D}[q_2(\boldsymbol{\Omega})|p(\boldsymbol{\Omega})] = \frac{1}{2} \sum_{n,j} \boldsymbol{Q}_{n,j}^2 l_n + \boldsymbol{R}_{n,j}^2 l_n - 1 - \log(\boldsymbol{Q}_{n,j}^2 l_n),$$
$$\mathcal{D}[q_3(\boldsymbol{W})|p(\boldsymbol{W})] = \frac{1}{2} \sum_{n,j} \boldsymbol{V}_{n,j}^2 + \boldsymbol{T}_{n,j}^2 - 1 - \log(\boldsymbol{V}_{n,j}^2).$$

By summation over the different modalities we finally obtain the total KL divergences.

Appendix C.

We provide in this Appendix details for the experiments on real data.

- The number of random features for the GP estimation was set to 10, as it was enough to recover the temporal sources in the synthetic experiments.
- The γ parameter controlling monotonicity was set to $\gamma_m = 10^7$ for each imaging modality ($F_m = 1,418,820$ imaging features and $N_m = 6$ sources) and $\gamma_c = 1$ for ADAS13 ($C_c = 1$ scalar feature).
- The lower bound was optimized using the ADAM optimizer [18].
- We used an alternate optimization scheme between the spatio-temporal parameters and the time-shift of [2000, 1000] iterations repeated 20 times, followed by 30000 iterations in which we only optimized the spatio-temporal parameters.
- The expectation terms in the lower bound were approximated using only one Monte-Carlo sample as proposed in [20].
- The table below gives the learning rates (LR) of all the parameters of the model.

Table 1: Learning rates (LR) of the different parameters of the model.

	$\boldsymbol{\theta}$	$\boldsymbol{M}$	$\boldsymbol{P}$	$\boldsymbol{Z}$	$\sigma,\ \nu$	$\boldsymbol{\delta}$
LR	10^{-2}	10^{-3}	10^{-1}	10^{-1}	10^{-2}	10^{-4}

Appendix D.

In this Appendix, we first provide a pseudo-code for sampling from a normal distribution using the reparameterization trick (see Algorithm 1). The second pseudo-code (Algorithm 2) details the steps to compute the lower bound $\boldsymbol{\mathcal{L}}_m$ for a given imaging modality m. We recall that we want to optimize the following sets of parameters (see Section 3.5): $\boldsymbol{\delta} = \{\delta_p\}_{p=0}^{P}$, $\boldsymbol{Z}$, $\sigma = \{\sigma_m\}_{m=1}^{M}$, $\nu = \{\nu_c\}_{c=1}^{C}$, $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^{M} \cup \{\theta_c\}_{c=1}^{C}$, and $\boldsymbol{\psi} = \{\psi_m\}_{m=1}^{M}$. Where P is the number of subjects, M the number of imaging modalities, C the number of scalar features, and N_m the number of spatio-temporal sources for a given modality m.

$$\begin{aligned} \boldsymbol{\theta} &= \{\boldsymbol{R}^m_{n:}, \boldsymbol{Q}^m_{n:}, \boldsymbol{T}^m_{n:}, \boldsymbol{V}^m_{n:}, l_n, n \in [1, N_m]\}_{m=1}^{M} \cup \{\boldsymbol{R}_{c:}, \boldsymbol{Q}_{c:}, \boldsymbol{T}_{c:}, \boldsymbol{V}_{c:}, l_c, \}_{c=1}^{C}, \\ \boldsymbol{\psi} &= \{\boldsymbol{M}^m_{n:}, \boldsymbol{P}^m_{n:}, n \in [1, N_m]\}_{m=1}^{M}. \end{aligned} \tag{1}$$

Similarly to Algorithm 2, we can derive a function LOSS_SCALAR when dealing with scalar scores by removing the computations on the spatial sources. Finally the last pseudo-code (Algorithm 3) details the model optimization. For sake of clarity we denote by $\boldsymbol{\Pi}$, the set of all the spatio-temporal parameters of the model.

Algorithm 1 Sampling from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the reparameterization trick.

function RT($\boldsymbol{\mu}, \boldsymbol{\Sigma}$)
  $\boldsymbol{\epsilon} \leftarrow$ random sample from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon}$ ▷ Gives one sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  Return $\boldsymbol{z}$
end function

Algorithm 2 Compute loss for a given imaging modality m.

function LOSS_IMAGE($\boldsymbol{Y}^m, \theta_m, \psi_m, \boldsymbol{Z}^m, \sigma_m, \boldsymbol{\delta}, \gamma_m, N_m, F_m, P$)
For ease of notation we drop the m index in the pseudo-code.
  for n=1 to N **do** ▷ For each source
    $\boldsymbol{B}_{n:} = \text{RT}(\boldsymbol{M}_{n:}, diag(\boldsymbol{P}_{n,:}))$ ▷ Sampling from q_1
    $\boldsymbol{\omega}^n = \text{RT}(\boldsymbol{R}_{n:}, diag(\boldsymbol{Q}^2_{n:}))$ ▷ Sampling from q_2
    $\boldsymbol{w}^n = \text{RT}(\boldsymbol{T}_{n:}, diag(\boldsymbol{V}^2_{n:}))$ ▷ Sampling from q_3
    $\boldsymbol{A}_{n:} = \boldsymbol{B}_{n:}\boldsymbol{\Sigma}^n$ ▷ Convolution of the sparse code of source n at a given spatial resolution
    $\boldsymbol{S}_{:n}(\boldsymbol{\delta}) = \phi(\boldsymbol{\delta}(\boldsymbol{\omega}^n)^T)\boldsymbol{w}^n$ ▷ Compute temporal trajectory of source n
    $\boldsymbol{S}'_{:n}(\boldsymbol{\delta}) = \frac{d\phi(\boldsymbol{\delta}(\boldsymbol{\omega}^n)^T)}{d\boldsymbol{\delta}}\boldsymbol{w}^n$ ▷ Compute derivative of temporal trajectory of source n
  end for
  $\boldsymbol{\Omega} \leftarrow$ block diagonal matrix containing all the set of $(\boldsymbol{\omega}^n)^T$
  $\boldsymbol{W} \leftarrow$ block diagonal matrix containing all the set of $\boldsymbol{w}^n$
  $\mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{\Omega},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\delta},\sigma))] \approx \sum_p -\frac{F}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\boldsymbol{Y}_{p:} - \boldsymbol{S}_{p:}\boldsymbol{A} - \boldsymbol{Z}_{p:}||^2$
  $\mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}|\boldsymbol{\Omega},\boldsymbol{W},\boldsymbol{\delta},\gamma))] \approx -\sum_{p,n}\log((1+\exp(-\gamma\boldsymbol{S}'_{p,n}(\boldsymbol{\delta})))$
    ▷ The two expectations terms are approximated using only one Monte-Carlo sample as proposed in [20].
  $\text{KL} = \mathcal{D}[q_1(\boldsymbol{B})||p(\boldsymbol{B})] + \mathcal{D}[q_2(\boldsymbol{\Omega})||p(\boldsymbol{\Omega})] + \mathcal{D}[q_3(\boldsymbol{W})||p(\boldsymbol{W})]$ ▷ This tern is computed using approximations and formulas of Appendix B.
  $\boldsymbol{\mathcal{L}} = \mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{\Omega},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\delta},\sigma))] + \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}|\boldsymbol{\Omega},\boldsymbol{W},\boldsymbol{\delta},\gamma))] - \text{KL}$
  Return $\boldsymbol{\mathcal{L}}$
end function

Algorithm 3 Model optimization.

function OPTIMIZE($\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\Pi}, \boldsymbol{\delta}$, n_iter0, n_iter1, n_iter2)
  Initialize $\boldsymbol{\Pi}^{(0)}$, $\boldsymbol{\delta}^{(0)}$
  $i, j, k = 0$
  while i $\leq$ n_iter0 **do**
    for l=1 to n_iter1 **do** ▷ Optimizing spatio-temporal parameters only
      $\boldsymbol{\mathcal{L}} = 0$
      for m=1 to M **do** ▷ For each modality
        $\boldsymbol{\mathcal{L}}$ += LOSS_IMAGE($\boldsymbol{Y}^m, \theta_m, \psi_m, \boldsymbol{Z}^m, \sigma_m, \boldsymbol{\delta}, \gamma_m, N_m, F_m, P$)
      end for
      for c=1 to C **do** ▷ For each scalar feature
        $\boldsymbol{\mathcal{L}}$ += LOSS_SCALAR($\boldsymbol{V}_{:c}, \theta_c, \nu_c, \boldsymbol{\delta}, \gamma_m, P$)
      end for
      Compute $\frac{d\boldsymbol{\mathcal{L}}}{d\boldsymbol{\Pi}^{(j)}}$ through backpropagation
      $\boldsymbol{\Pi}^{(j+1)} = \text{ADAM}(\frac{d\boldsymbol{\mathcal{L}}}{d\boldsymbol{\Pi}^{(j)}}, \boldsymbol{\Pi}^{(j)}, \text{LR}(\boldsymbol{\Pi}))$ ▷ The spatio-temporal parameters are optimized by gradient descent using the ADAM optimizer. LR refers to the overall set of learning rates (cf Appendix C.)
      j += 1
    end for
    for l=1 to n_iter2 **do** ▷ Optimizing time-shift only
      $\boldsymbol{\mathcal{L}} = 0$
      for m=1 to M **do**
        $\boldsymbol{\mathcal{L}}$ += LOSS_IMAGE($\boldsymbol{Y}^m, \theta_m, \psi_m, \boldsymbol{Z}^m, \sigma_m, \boldsymbol{\delta}, \gamma_c, N_m, F_m, P$)
      end for
      for c=1 to C **do**
        $\boldsymbol{\mathcal{L}}$ += LOSS_SCALAR($\boldsymbol{V}_{:c}, \theta_c, \nu_c, \boldsymbol{\delta}, \gamma_c, P$)
      end for
      Compute $\frac{d\boldsymbol{\mathcal{L}}}{d\boldsymbol{\delta}^{(k)}}$ through backpropagation
      $\boldsymbol{\delta}^{(k+1)} = \text{ADAM}(\frac{d\boldsymbol{\mathcal{L}}}{d\boldsymbol{\delta}^{(k)}}, \boldsymbol{\delta}^{(k)}, \text{LR}(\boldsymbol{\delta}))$
      k += 1
    end for
    i += 1
  end while
end function

Appendix E.

In this Appendix, we show results obtained with standard methods (ICA, NMF, PCA) when applied within the experimental setting of Section 4.2. We recall that for these experiments observations were randomly aligned along the time-axis. The goal was to assess the ability of the different methods to reconstruct the spatio-temporal sources underlying the data when the time-axis is unknown. Results obtained in Table 1 show a substantial decrease of performances for the MSE and SSIM compared to MGPA (cf Table 2 in Section 4.2). Indeed, these methods do not consider time as a variable on which inference is required, thus preventing them from reconstructing correctly the temporal sources. Figure 1 shows an example of reconstruction when using ICA. We observe that even though the spatial reconstruction remains acceptable, the estimated temporal sources are not interpretable as ICA reconstructs the data using the time-axis on which observations have been mixed.

Table 1: MSE and SSIM between respectively the ground truth temporal and spatial sources with respect to the ones estimated by the different standard methods.

	Temporal (MSE)	Spatial (SSIM)
ICA	0.24 ± 0.08	$54\% \pm 2$
NMF	0.25 ± 0.03	$22\% \pm 14$
PCA	0.66 ± 0.05	$9\% \pm 3$

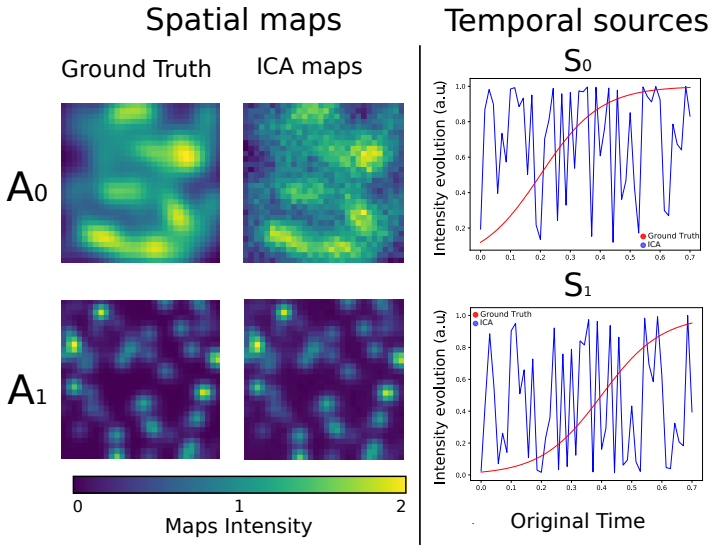


Figure 1: Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm), the maps estimated by ICA. Temporal sources: Ground truth temporal sources (red) along with sources estimated by ICA (blue).

Appendix F.

We provide in this Appendix details on the model convergence when applied on the ADNI data. The training was divided in three iterations of 30000 epochs each. During the two first iterations the spatio-temporal parameters and the time-shift are trained alternatively following a scheme of [2000,1000] epochs ten times. The third iteration only optimizes the spatio-temporal parameters. In Figure 1, we show the evolution of the total loss and the different terms composing it during training. The term reconstruction cost stands for $\sum_m \mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}^m|\boldsymbol{B}^m,\boldsymbol{\Omega}^m,\boldsymbol{W}^m,\boldsymbol{Z}^m,\boldsymbol{\delta},\sigma_m))]$, monotonicity cost for $\sum_m \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^m|\boldsymbol{\Omega}^m,\boldsymbol{W}^m,\boldsymbol{\delta},\gamma_m))]$ and KL for $\sum_m \mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)] + \mathcal{D}[q_2(\boldsymbol{\Omega}^m)||p(\boldsymbol{\Omega}^m)] + \mathcal{D}[q_3(\boldsymbol{W}^m)||p(\boldsymbol{W}^m)]$. We observe that through the first two iterations the reconstruction and monotonicity costs decrease, and become stable during the last iteration. Differently, the KL cost increases during the first iteration as the model is driven by the reconstruction and monotonicity constraints. The KL term decreases during the second iteration, thus regularizing the model, before becoming stable during the third iteration. We also note that the graphs in Figure 1 show convergence profiles typical of those obtained with stochastic variational inference schemes, such as with Variational Autoencoders or Bayesian Neural Networks.

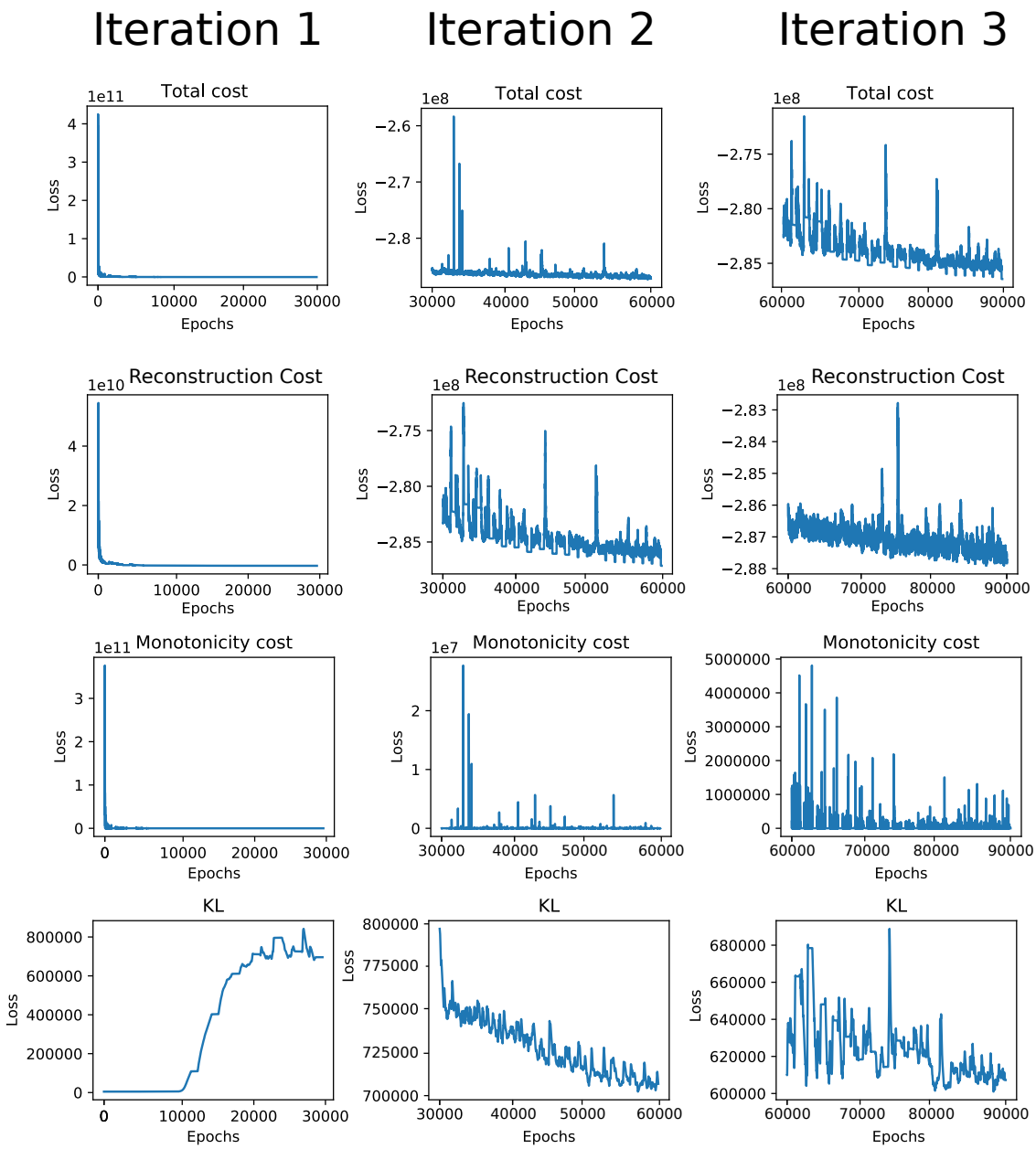


Figure 1: Evolution of the total loss, reconstruction cost, monotonicity cost and KL during training. Each iteration corresponds to 30000 epochs.

Appendix G.

In this Appendix, we provide the results obtained when applying ICA, NMF and PCA on the ADNI data of Section 4.3.1. We used the three imaging modalities for each subject and concatenated these images in a (544×4256460) matrix. Our goal was to compare the spatio-temporal processes extracted using these standard methods with the ones from MGPA. We recall that in the case of MGPA the model automatically re-aligns the observations following monotonic assumptions for each biomarker, while these standard methods don't perform any inference on the time variable. Therefore, we created three experimental settings in which we changed the observations' alignment. In the first one, subjects were aligned by their chronological age (Figures 1, 2 and 3), in the second one by ADAS13 (Figures 4, 5 and 6) and in the last one time was randomly initialized like in the experiments of Section 4.3.3 (Figures 7, 8 and 9). We extracted six spatio-temporal sources for each method and each time-alignment, like in 4.3.2.

We observe that the temporal profiles are generally noisy and hard to interpret due to the lack of constraints on the temporal evolution. This motivates the need of smooth and monotonic constraints as in MGPA. Moreover, due to the concatenation of all the modalities they all share the same temporal patterns. This is an important difference with the modality-specific modelling of MGPA. Finally, we note that the spatial patterns associated with these methods are very similar, independently from the time-initialization, while the temporal sources substantially differ. This is also true when time is randomly initialized. These observations point to the challenge of giving a clinical interpretation of the results obtained with these approaches, and therefore to the need of plausible spatio-temporal constraints as provided in MGPA.

Subjects aligned by age.

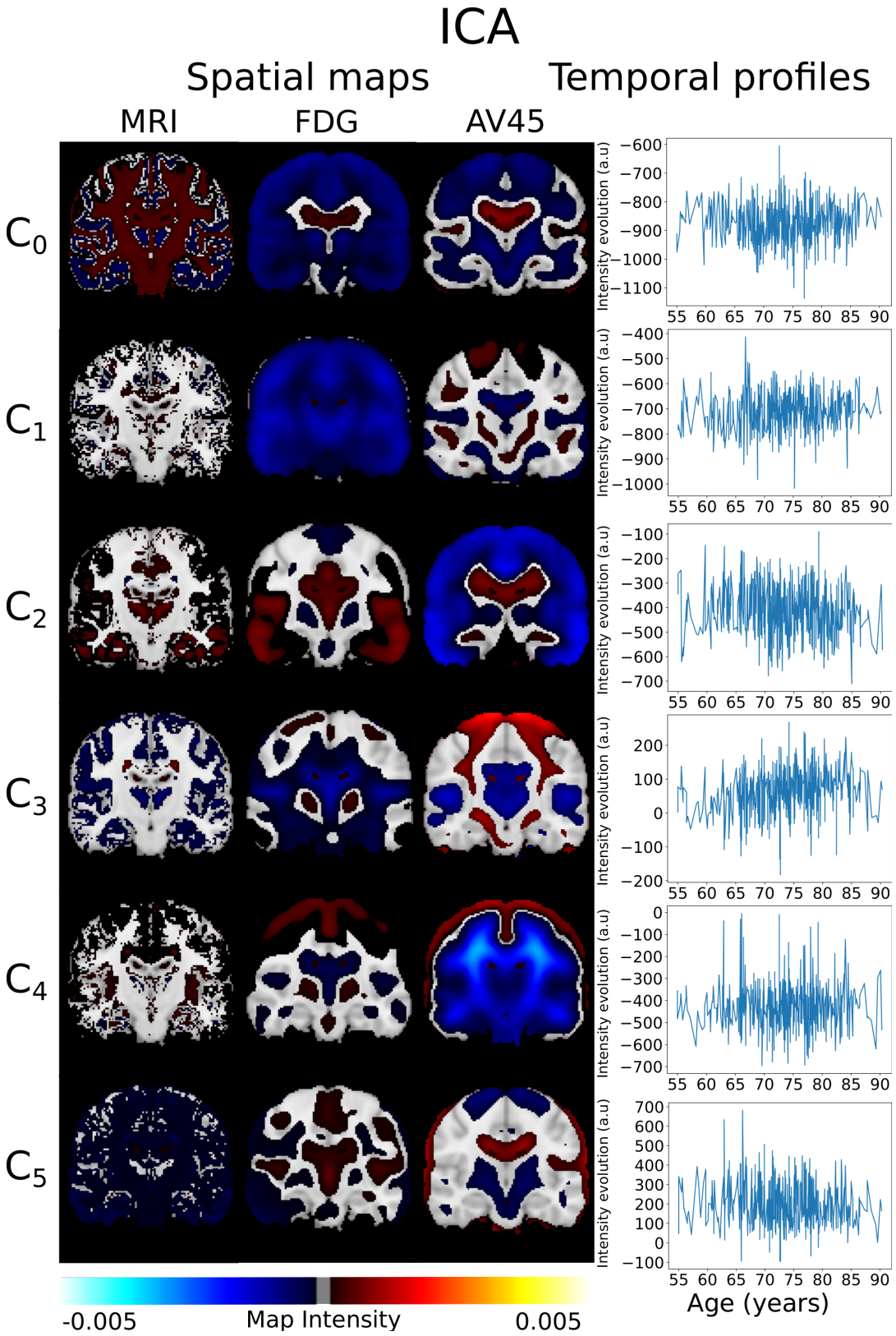


Figure 1: Spatio-temporal processes extracted by ICA with subjects aligned by age.

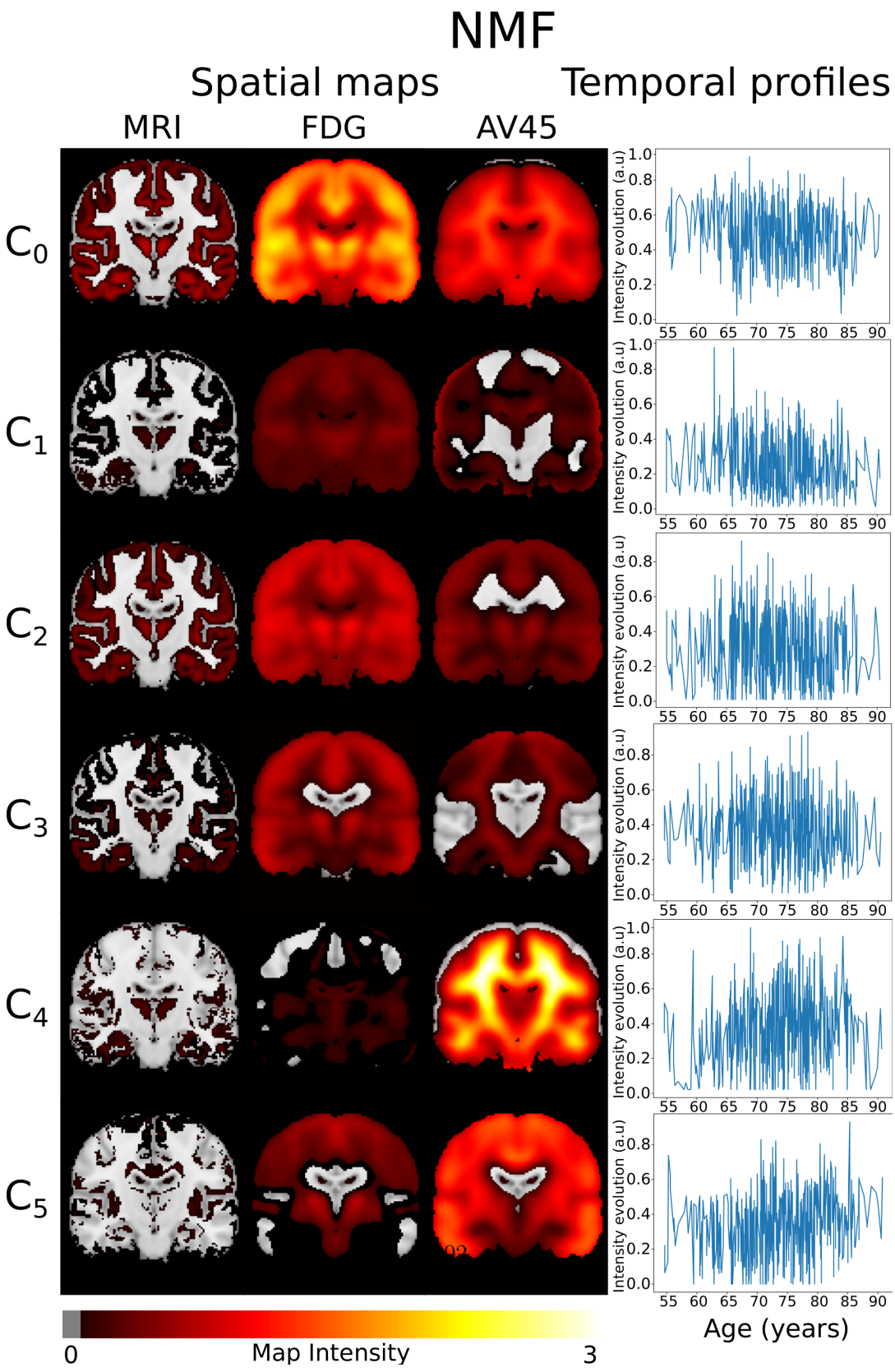


Figure 2: Spatio-temporal processes extracted by NMF with subjects aligned by age.

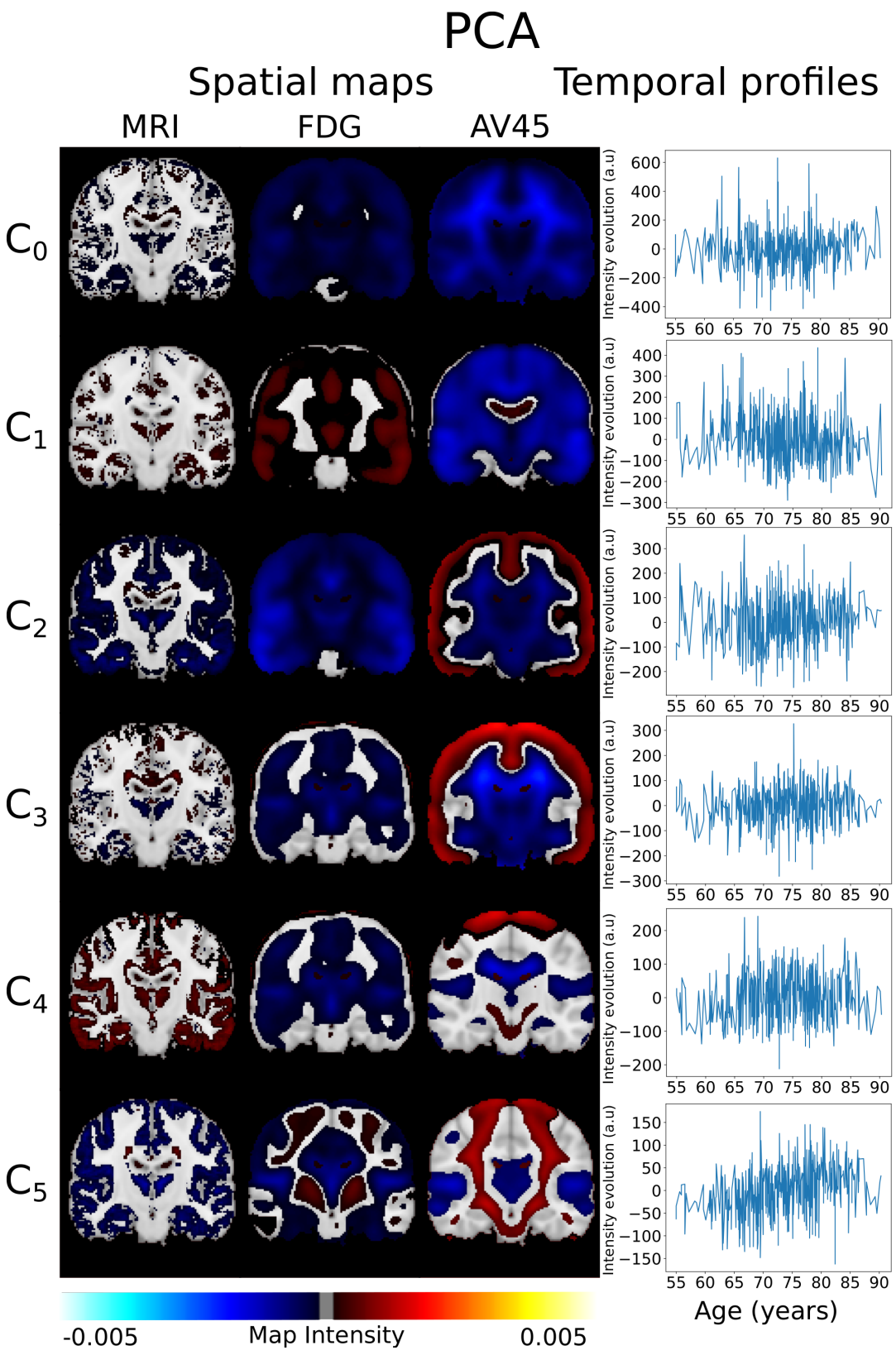


Figure 3: Spatio-temporal processes extracted by PCA with subjects aligned by age.

Subjects aligned by ADAS13.

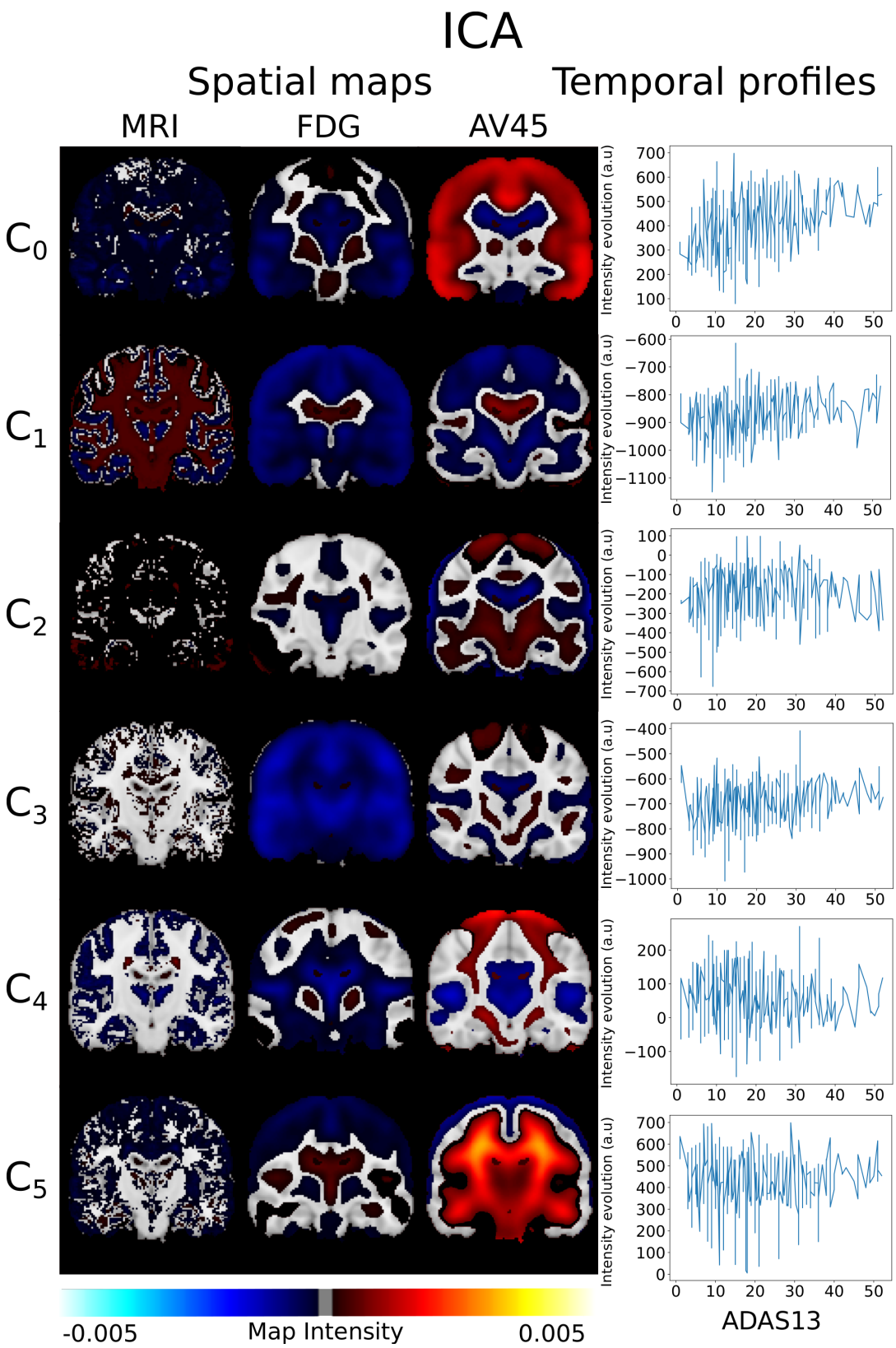


Figure 4: Spatio-temporal processes extracted by ICA with subjects aligned by ADAS13.

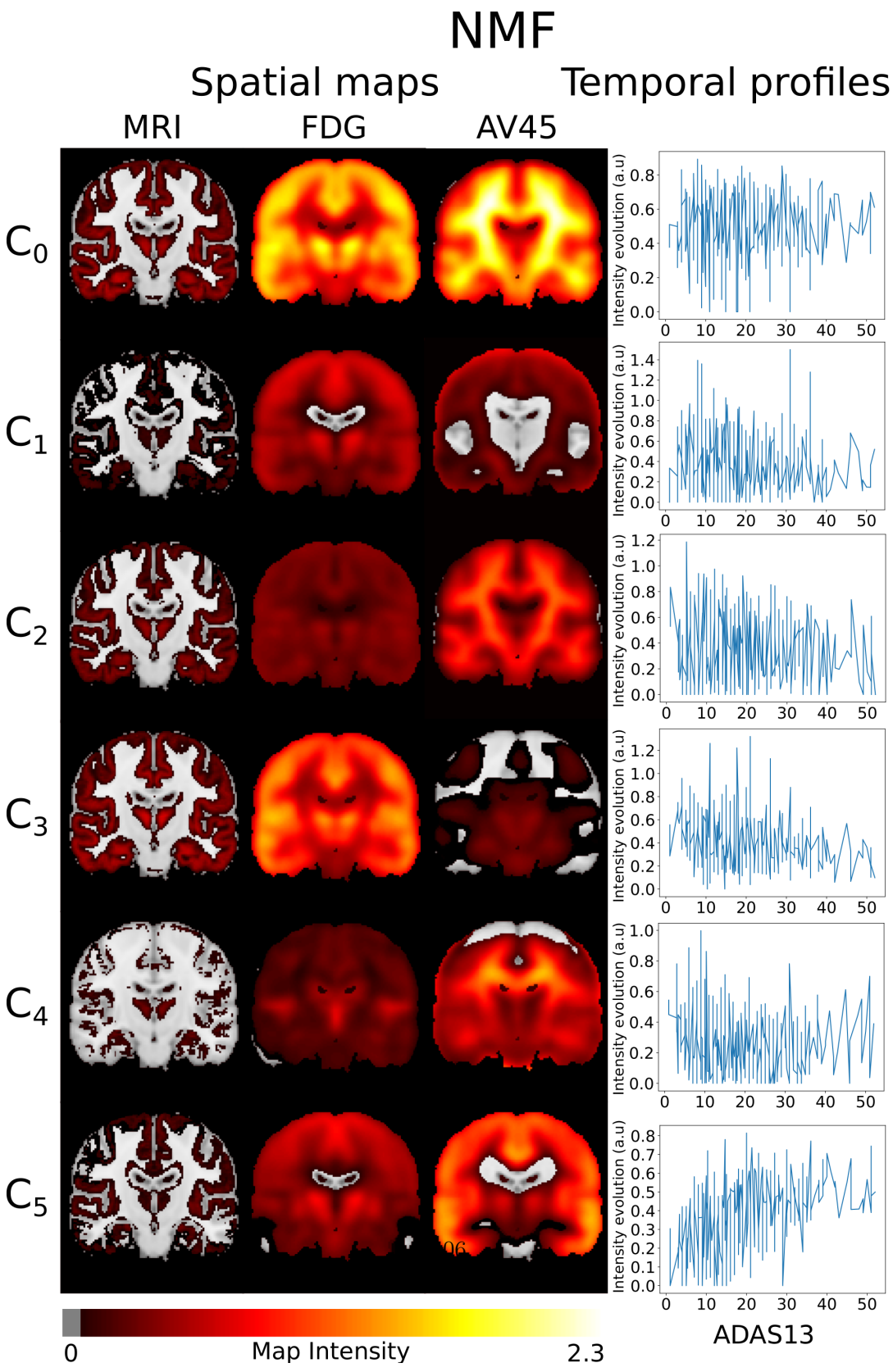


Figure 5: Spatio-temporal processes extracted by NMF with subjects aligned by ADAS13.

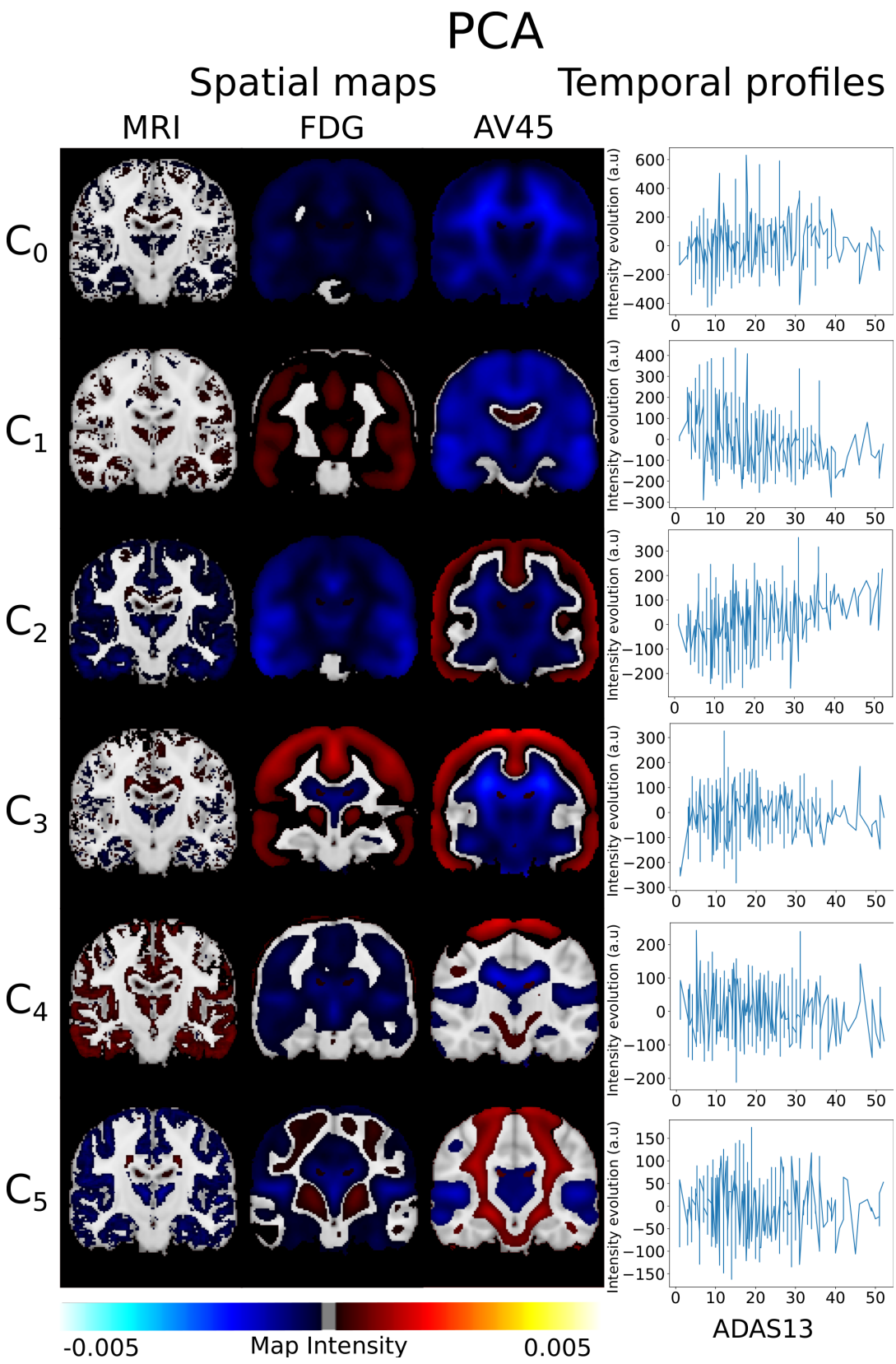


Figure 6: Spatio-temporal processes extracted by PCA with subjects aligned by ADAS13.

Subjects randomly aligned.

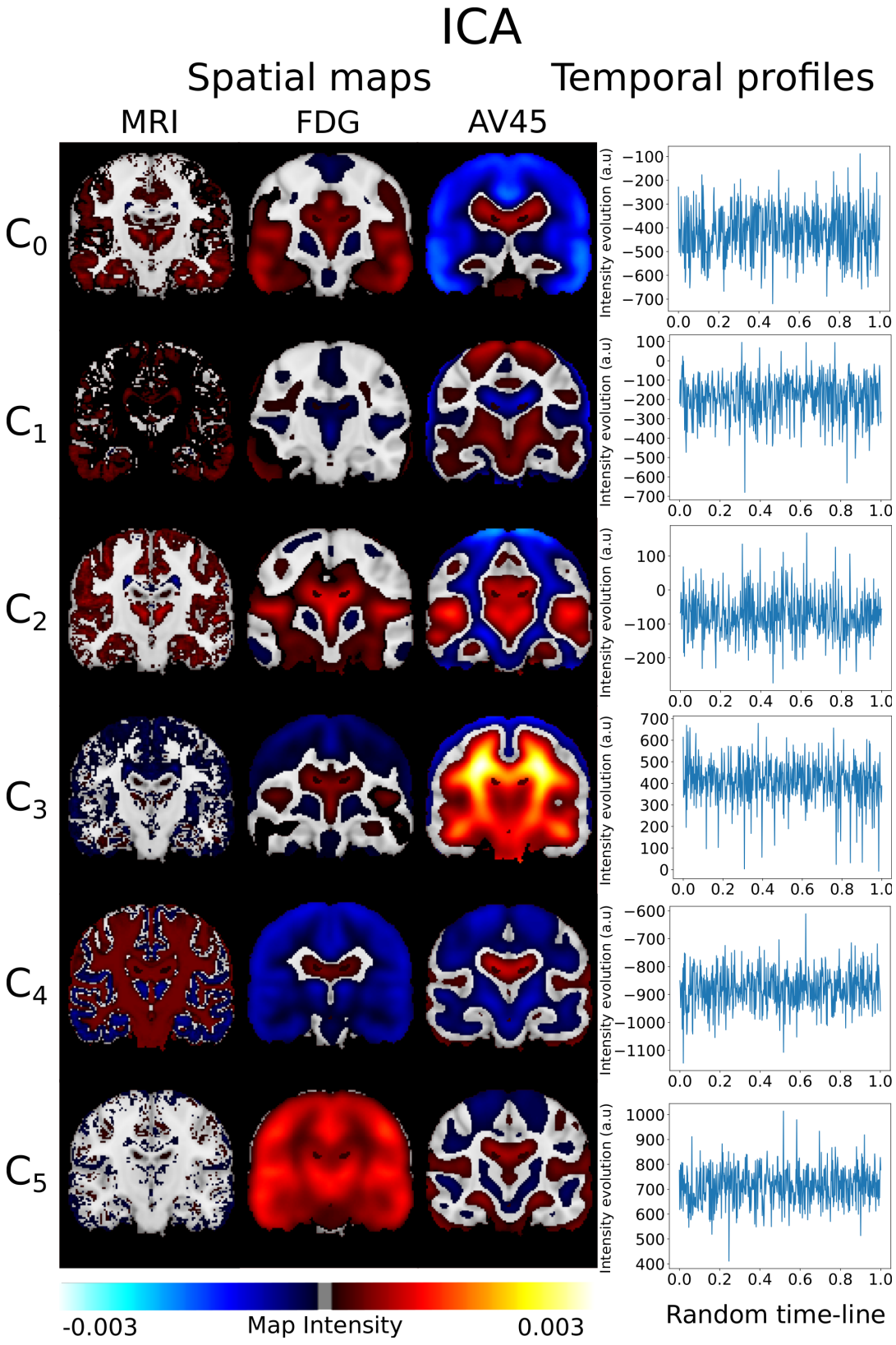


Figure 7: Spatio-temporal processes extracted by ICA with subjects randomly aligned.

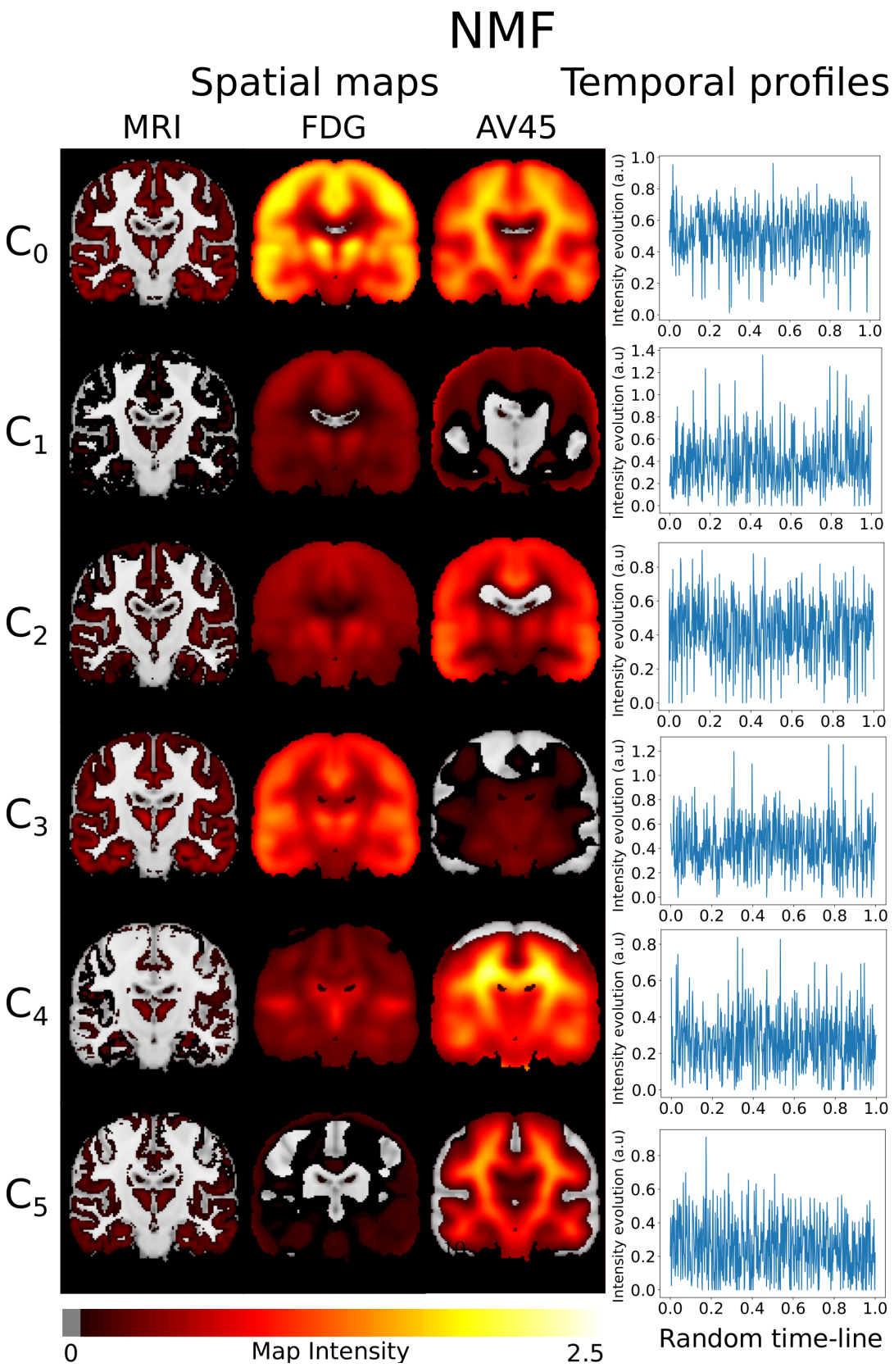


Figure 8: Spatio-temporal processes extracted by NMF with subjects randomly aligned.

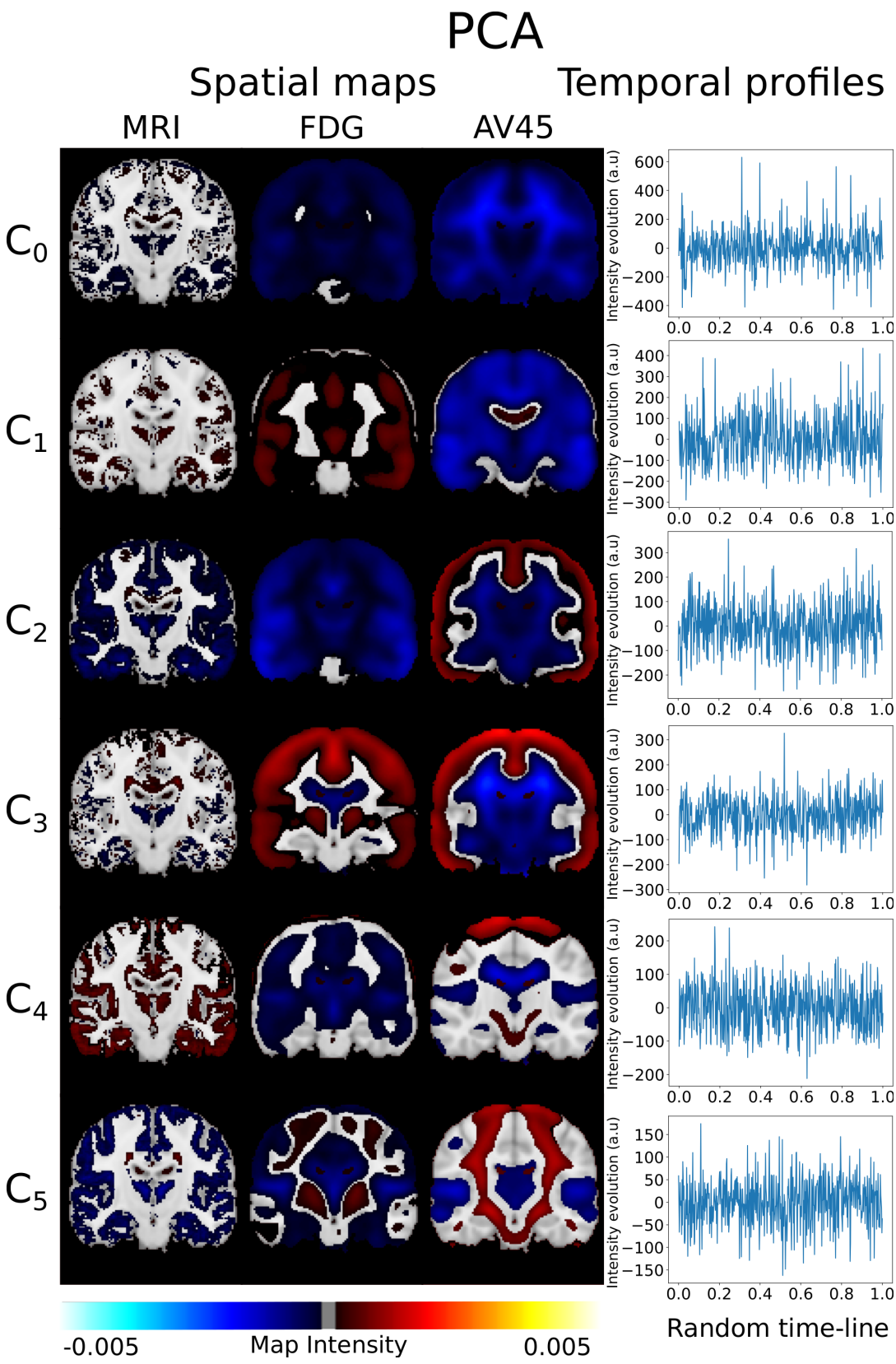


Figure 9: Spatio-temporal processes extracted by PCA with subjects randomly aligned.

Constraining the Dynamics of Deep Probabilistic Models

Marco Lorenzi[1] and Maurizio Filippone[1]

1. Université Côte d'Azur, INRIA Sophia Antipolis, EPIONE research group, France

2. EURECOM, Department of Data Science, Sophia Antipolis, France

Originally published in:

International Conference on Machine Learning (ICML), PMLR 80:3233-3242, 2018

Abstract

We introduce a novel generative formulation of deep probabilistic models implementing "soft" constraints on their function dynamics. In particular, we develop a flexible methodological framework where the modeled functions and derivatives of a given order are subject to inequality or equality constraints. We then characterize the posterior distribution over model and constraint parameters through stochastic variational inference. As a result, the proposed approach allows for accurate and scalable uncertainty quantification on the predictions and on all parameters. We demonstrate the application of equality constraints in the challenging problem of parameter inference in ordinary differential equation models, while we showcase the application of inequality constraints on the problem of monotonic regression of count data. The proposed approach is extensively tested in several experimental settings, leading to highly competitive results in challenging modeling applications, while offering high expressiveness, flexibility and scalability.

1 Introduction

Modern machine learning methods have demonstrated state-of-art performance in representing complex functions in a variety of applications. Nevertheless, the translation of complex learning methods in natural sciences and in the clinical domain is still challenged by the need of interpretable solutions. To this end, several approaches have been proposed in order to constrain the solution dynamics to plausible forms such as boundedness Da Veiga & Marrel (2012), monotonicity Riihimäki & Vehtari (2010), or mechanistic behaviors Alvarez et al. (2013). This is a crucial requirement to provide a more precise and realistic description of natural phenomena. For example, monotonicity of the interpolating function is a common assumption when modeling disease progression in neurodegenerative diseases Lorenzi et al. (2017); Donohue et al. (2014), while bio-physical or mechanistic models are necessary when analyzing and simulating experimental data in bio-engineering Vyshemirsky & Girolami (2007); Konukoglu et al. (2011).

However, accounting for the complex properties of biological systems in data-driven modeling approaches poses important challenges. For example, functions are often non-smooth and characterized by nonstationaries which are difficult to encode in "shallow" models. Complex cases can arise already in classical ODE systems for certain configurations of the parameters, where functions can exhibit sudden temporal changes Goel et al. (1971); FitzHugh (1955). Within this context,

approaches based on stationary models, even when relaxing the smoothness assumptions, may lead to suboptimal results for both data modeling (interpolation), and estimation of dynamics parameters. To provide insightful illustrations of this problem we anticipate the results of Section 4.4.1 and Figure 5. Moreover, the application to real data requires to account for the uncertainty of measurements and underlying model parameters, as well as for the – often large – dimensionality characterizing the experimental data. Within this context, deep probabilistic approaches may represent a promising modeling tool, as they combine the flexibility of deep models with a systematic way to reason about uncertainty in model parameters and predictions. The flexibility of these approaches stems from the fact that deep models implement compositions of functions, which considerably extend the complexity of signals that can be represented with "shallow" models LeCun et al. (2015). Meanwhile, their probabilistic formulation introduces a principled approach to quantify uncertainty in parameters estimation and predictions, as well as to model selection problems Neal (1996); Ghahramani (2015).

In this work, we aim at extending deep probabilistic models to account for constraints on their dynamics. In particular, we focus on a general and flexible formulation capable of imposing a rich set of constraints on functions and derivatives of any order. We focus on: i) *equality constraints* on the function and its derivatives, required when the model should satisfy given physical laws implemented through a mechanistic description of a system of interest; and ii) *inequality constraints*, arising in problems where the class of suitable functions is characterized by specific properties, such as monotonicity or convexity/concavity (Riihimäki & Vehtari, 2010).

In case of equality constraints, we tackle the challenge of parameters inference in Ordinary Differential Equations (ODE). Exact parameter inference of ODE models is computationally expensive due to the need for repeatedly solving ODEs within the Bayesian setting. To this end, previous works attempted to recover tractability by introducing approximate solutions of ODEs (see, e.g., Macdonald & Husmeier (2015) for a review). Following these ideas, we introduce "soft" constraints through a probabilistic formulation that penalizes functions violating the ODE on a set of virtual inputs. Note that this is in contrast to previous approaches, such as the ones proposed with probabilistic ODE solvers Wheeler et al. (2014); Schober et al. (2014), where a given dynamics is strictly enforced to the model posterior. By deriving a lower bound on the model evidence, we enable the use of stochastic variational inference to achieve end-to-end posterior inference over model and constraint parameters.

In what follows we shall focus on a class of deep probabilistic models implementing a composition of Gaussian processes (GPs) (Rasmussen & Williams, 2006) into Deep Gaussian Processes (DGPs) (Damianou & Lawrence, 2013). More generally, our formulation can be straightforwardly extended to probabilistic Deep Neural Networks (DNNs) (Neal, 1996). On the practical side, our formulation allows us to take advantage of automatic differentiation tools, leading to flexible and easy-to-implement methods for inference in constrained deep probabilistic models. As a result, our method scales linearly with the number of observations and constraints. Furthermore, in the case of mean-field variational inference, it also scales linearly with the number of parameters in the constraints. Finally, it can easily be parallelized/distributed and exploit GPU computing.

Through an in-depth series of experiments, we demonstrate that our proposal achieves state-of-the-art performance in a number of constrained modeling problems while being characterized by attractive scalability properties. The paper is organized as follows: Section 2 reports on related work, whereas the core of the methodology is presented in Section 3. Section 4 contains an in-depth validation of the proposed model against the state-of-the-art. We demonstrate the application of equality constraints in the challenging problem of parameter inference in ODE, while we showcase

the application of inequality constraints in the monotonic regression of count data. Additional insights and conclusions are given in Section 5. Results that we could not fit in the manuscript are deferred to the supplementary material.

2 Related Work

Equality constraints where functions are enforced to model the solution of ODE systems have been considered in a variety of problems, particularly in the challenging task of accelerated inference of ODE parameters. Previous approaches to accelerate ODE parameter optimization involving interpolation date back to Varah (1982). This idea has been developed in several ways, including splines, GPs, and Reproducing Kernel Hilbert spaces. Works that employ GPs as interpolants have been proposed in Ramsay et al. (2007), Liang & Wu (2008), Calderhead et al. (2009), and Campbell & Steele (2012). Such approaches have been extended to introduce a novel formulation to regularize the interpolant based on the ODE system, notably Dondelinger et al. (2013); Barber & Wang (2014). An in-depth analysis of the model in Barber & Wang (2014) is provided by Macdonald et al. (2015). Recently, Gorbach et al. (2017) extended previous works by proposing mean-field variational inference to obtain an approximate posterior over ODE parameters. Our work improves previous approaches by considering a more general class of interpolants than "shallow" GPs, and proposes a scalable framework for inferring the family of interpolating functions jointly with the parameters of the constraint, namely ODE parameters.

Another line of research that builds on gradient matching approaches uses a Reproducing Kernel Hilbert space formulation. For example, González et al. (2014) proposes to exploit the linear part of ODEs to accelerate the interpolation, while Niu et al. (2016) exploits the quadratic dependency of the objective with respect to the parameters of the interpolant to improve the computational efficiency of the ODE regularization. Interestingly, inspired by Calandra et al. (2016), the latter approach was extended to handle nonstationarity in the interpolation through warping (Niu et al., 2018). The underlying idea is to estimate a transformation of the input domain to account for nonstationarity of the signal, in order to improve the fitting of stationary GP interpolants. A key limitation of this approach is the lack of a probabilistic formulation, which prevents one from approximating the posterior over ODE parameters. Moreover, the warping approach is tailored to periodic functions, thus limiting the generalization to more complex signals. In our work, we considerably improve on these aspects by effectively modeling the warping through GPs/DGPs that we infer jointly with ODE parameters.

Inequality constraints on the function derivatives have been considered in several works such as in Meyer (2008); Groeneboom & Jongbloed (2014); Mašić et al. (2017); Riihimäki & Vehtari (2010); Da Veiga & Marrel (2012); Salzmann & Urtasun (2010). In particular, the GP setting provides a solid and elegant theoretical background for tackling this problem; thanks to the linearity of differentiation, both mean and covariance functions of high-order derivatives of GPs can be expressed in closed form, leading to exact formulations for linearly-constrained GPs (Da Veiga & Marrel, 2012). In case of inequality constraints on the derivatives, instead, this introduces non-conjugacy between the likelihood imposing the derivative constraint and the GP prior, thus requiring approximations (Riihimäki & Vehtari, 2010). Although this problem can be tackled through sampling schemes or variational inference methods, such as Expectation Propagation (Minka, 2001), scalability to large dimensions and sample size represents a critical limitation. In this work, we extend these methods by considering a more general class of functions based on DGPs, and develop scalable inference that

makes our method applicable to large data and dimensions.

3 Methods

3.1 Equality constraints in probabilistic modeling

In this section we provide a derivation of the posterior distribution of our model when we introduce equality constraints in the dynamics. Let Y be a set of n observed multivariate variables $\mathbf{y}_i \in \mathbb{R}^s$ associated with measuring times t collected into $\mathbf{t}$; the extension where the n variables are measured at different times is notationally heavier but straightforward. Let $\mathbf{f}(t)$ be a multivariate interpolating function with associated noise parameters $\boldsymbol{\psi}$, and define F similarly to Y to be the realization of $\mathbf{f}$ at $\mathbf{t}$. In this work, $\mathbf{f}(t)$ will be either modeled using a GP, or deep probabilistic models based on DGPs. We introduce functional constraints on the dynamics of the components of $\mathbf{f}(t)$ by specifying a family of admissible functions whose derivatives of order h evaluated at the inputs $\mathbf{t}$ satisfy some given constraint

$$\mathcal{C}_{hi} = \left\{ \mathbf{f}(t) \middle| \frac{d^h f_i(\mathbf{t})}{dt^h} = \mathcal{H}_{hi}\left(t, \mathbf{f}, \frac{d\mathbf{f}}{dt}, \ldots, \frac{d^q \mathbf{f}}{dt^q}, \boldsymbol{\theta}\right)\Big|_{\mathbf{t}} \right\}.$$

Here the constraint is expressed as a function of the input, the function itself, and high-order derivatives up to order q. The constraint also includes $\boldsymbol{\theta}$ as dynamics parameters that should be inferred. We are going to consider the intersection of all the constraints for a set of indices $\mathcal{I}$ comprising pairs (h, i) of interest

$$\mathcal{C} = \bigcap_{(h,i)\in\mathcal{I}} \mathcal{C}_{hi}$$

To keep the notation uncluttered, and without loss of generality, in the following we will assume that all the terms are evaluated at $\mathbf{t}$; we can easily relax this by allowing for the constraints to be evaluated at different sampling points than $\mathbf{t}$. As a concrete example, consider the constraints induced by the Lotka-Volterra ODE system (more details in the experiments section); for this system, $\boldsymbol{\theta} = \{\alpha, \beta, \gamma, \delta\}$, and the family of functions is identified by the conditions

$$\left.\frac{dg_1(t)}{dt}\right|_{\mathbf{t}} = \left.\mathcal{H}_{11}\left(\mathbf{f}(t)\right)\right|_{\mathbf{t}} = \alpha f_1(\mathbf{t}) - \beta f_1(\mathbf{t}) f_2(\mathbf{t}),$$

$$\left.\frac{dg_2(t)}{dt}\right|_{\mathbf{t}} = \left.\mathcal{H}_{12}\left(\mathbf{f}(t)\right)\right|_{\mathbf{t}} = -\gamma f_2(\mathbf{t}) + \delta f_1(\mathbf{t}) f_2(\mathbf{t}),$$

where the products $f_1(\mathbf{t})f_2(\mathbf{t})$ are element-wise.

Denote by $\tilde{F} = \{\mathbf{f}_{hi}\}$ the set of realizations of $\mathbf{f}$ and of its derivatives at any required order h evaluated at timed $\mathbf{t}$. We define the constrained regression problem through two complementary likelihood-based elements: a data attachment term $p(Y|F, \boldsymbol{\psi})$, and a term quantifying the constraint on the dynamics, $p(\mathcal{C}|\tilde{F}, \boldsymbol{\theta}, \boldsymbol{\psi}_D)$, where $\boldsymbol{\psi}_D$ is the associated noise parameter. To solve the inference problem, we shall determine a lower bound for the marginal

$$p(Y, \mathcal{C}|\mathbf{t}, \boldsymbol{\psi}, \boldsymbol{\psi}_D) = \tag{1}$$
$$\int p(Y|F, \boldsymbol{\psi}) p(\mathcal{C}|\tilde{F}, \boldsymbol{\theta}, \boldsymbol{\psi}_D) p(F, \tilde{F}|\mathbf{t}, \boldsymbol{\psi}) p(\boldsymbol{\theta}) dF d\tilde{F} d\boldsymbol{\theta},$$

where

$$p(F, \tilde{F}|\mathbf{t}, \boldsymbol{\psi}) = p(\tilde{F}|F)p(F|\mathbf{t}, \boldsymbol{\psi}).$$

Note that $\tilde{F}$ is in fact completely identified by F.

Equation (1) requires specifying suitable models for both likelihood and functional constraints. This problem thus implies the definition of noise models for both observations and model dynamics. In the case of continuous observations, the likelihood can be assumed to be Gaussian:

$$p(Y|F, \boldsymbol{\psi}) = \mathcal{N}(Y|F, \Sigma(\boldsymbol{\psi})), \tag{2}$$

where $\Sigma(\boldsymbol{\psi})$ is a suitable multivariate covariance. Extensions to other likelihood functions are possible, and in the experiments we show an application to regression on counts where the likelihood is Poisson with rates equal to the exponential of the elements of F.

Concerning the noise model for the derivative observations, we assume independence across the constraints $\mathcal{C}_{hi}$ so that

$$p(\mathcal{C}|\tilde{F}, \theta, \boldsymbol{\psi}_D) = \prod_{(h,i)\in\mathcal{I}} p(\mathcal{C}_{hi}|\tilde{F}, \theta, \boldsymbol{\psi}_D). \tag{3}$$

We can again assume a Gaussian likelihood:

$$p(\mathcal{C}_{hi}|\tilde{F}, \theta, \boldsymbol{\psi}_D) = \prod_t \mathcal{N}(\mathbf{f}_{hi}(t)|\mathcal{H}_{hi}(t, \tilde{F}, \theta), \boldsymbol{\psi}_D), \tag{4}$$

or, in order to account for potentially heavy-tailed error terms on the derivative constraints, we can assume a Student-t distribution:

$$p(\mathcal{C}_{hi}|\tilde{F}, \theta, \boldsymbol{\psi}_D) = \prod_t \mathcal{T}(\mathbf{f}_{hi}(t)|\mathcal{H}_{hi}(t, \tilde{F}, \theta), \boldsymbol{\psi}_D, \nu), \tag{5}$$

where $\mathcal{T}(z|\mu, \lambda, \nu) \propto \frac{1}{\lambda}[1 + \frac{(z-\mu)^2}{\nu\lambda^2}]^{-(\nu+1)/2}$. We test these two noise models for $\tilde{F}$ in the experiments.

3.2 Inequality constraints in probabilistic modeling

In the case of inequality constraints we can proceed analogously as in the previous section. In particular, we are interested in the class of functions satisfying the following conditions:

$$\mathcal{C}_{hi} = \left\{ \mathbf{f}(t) \middle| \frac{d^h f_i(\mathbf{t})}{dt^h} > \mathcal{H}_{hi}\left(t, \mathbf{f}, \frac{d\mathbf{f}}{dt}, \ldots, \frac{d^q\mathbf{f}}{dt^q}, \boldsymbol{\theta}\right)\Big|_{\mathbf{t}} \right\}.$$

For example, a monotonic univariate regression problem can be obtained with a constraint of the form $\frac{df}{dt} > 0$.

In this case, the model dynamics can be enforced by a logistic function:

$$p(\mathcal{C}_{hi}|\tilde{F}, \psi_D) = \prod_{j=1}^{n} \frac{1}{1 + \exp(-\psi_D \frac{df}{dt}(t_j))}, \tag{6}$$

where the parameter ψ_D controls the strength of the monotonicity constraint.

3.3 Optimization and inference in constrained regression with dgps

After recalling the necessary methodological background, in this section we derive an efficient inference scheme for the model posterior introduced in Section 3.1.

To recover tractability, our scheme leverages on recent advances in modeling and inference in DGPs through approximation via random feature expansions (Rahimi & Recht, 2008; Cutajar et al., 2017). Denoting with $F^{(l)}$ the GP random variables at layer l, an (approximate) DGP is obtained by composing GPs approximated by Bayesian linear models, $F^{(l)} \approx \Phi^{(l)} W^{(l)}$. The so-called random features $\Phi^{(l)}$ are obtained by multiplying the layer input by a random matrix $\Omega^{(l)}$ and by applying a nonlinear transformation $h(\cdot)$. For example, in case of the standard RBF covariance, the elements in $\Omega^{(l)}$ are Gaussian distributed with covariance function parameterized through the length-scale of the RBF covariance. The nonlinearity is obtained through trigonometric functions, $\mathbf{h}(\cdot) = (\cos(\cdot), \sin(\cdot))$, while the prior over the elements of $W^{(l)}$ is standard normal. As a result, the interpolant becomes a Bayesian Deep Neural Network (DNN), where for each layer we have weights $\Omega^{(l)}$ and $W^{(l)}$, and activation functions $\mathbf{h}(\cdot)$ applied to the input of each layer multiplied by the weights $\Omega^{(l)}$.

3.3.1 Derivatives in dgps with random feature expansions

To account for function derivatives consistently with the theory developed in Cutajar et al. (2017), we need to extend the random feature expansion formulation of DGPs to high-order derivatives. Fortunately, this is possible thanks to the chain rule and to the closure under linear operations of the approximated GPs. More precisely, the derivatives of a "shallow" GP model with form $F = \mathbf{h}(\mathbf{t}\Omega)W$ can still be expressed through linear composition of matrix-valued operators depending on W and Ω only: $\frac{dF}{dt} = \frac{d\mathbf{h}(\mathbf{t}\Omega)}{dt} W$. The computational tractability is thus preserved and the GP function and derivatives are identified by the same sets of weights Ω and W. The same principle clearly extends to DGP architectures where the derivatives at each layer can be combined following the chain rule to obtain the derivatives of the output function with respect to the input.

3.3.2 Variational lower bound

In the constrained DGP setting, we are interested in carrying out inference of the functions $F^{(l)}$ and of the associated covariance parameters at all layers. Moreover, we may want to infer any dynamics parameters $\boldsymbol{\theta}$ that parameterize the constraint on the derivatives. Within this setting, the inference of the latent variables $F^{(l)}$ in the marginal (1) is generally not tractable. Nevertheless, the Bayesian DNN structure provided by the random feature approximation allows the efficient estimation of its parameters, and the tractability of the inference is thus recovered.

In particular, let $\boldsymbol{\Omega}$, $\mathbf{W}$, and $\boldsymbol{\psi}$ be the collections of all $\Omega^{(l)}$, $W^{(l)}$, and covariance and likelihood parameters, respectively. Recalling that we can obtain random features at each layer by sampling the elements in $\boldsymbol{\Omega}$ from a given prior distribution, we propose to tackle the inference problem through *variational inference* of the parameters $\mathbf{W}$ and $\boldsymbol{\theta}$. We could also attempt to infer $\boldsymbol{\Omega}$, although in this work we are going to assume them sampled from the prior with fixed randomness, which allows us to optimize covariance parameters using the reparameterization trick (option PRIOR-FIXED in Cutajar et al. (2017)). We also note that we could infer, rather than optimize, $\boldsymbol{\psi}$; we leave this for future work.

Using Jensen's inequality, the variational approach allows us to obtain a lower bound on the

log-marginal likelihood $\mathcal{L} := \log\left[p(Y, \mathcal{C}|\mathbf{t}, \boldsymbol{\Omega}, \boldsymbol{\psi}, \boldsymbol{\psi}_D)\right]$ of equation (1), as follows:

$$\begin{aligned}\mathcal{L} &\geq E_{q(\mathbf{W})}\left(\log[p(Y|\boldsymbol{\Omega}, \mathbf{W}, \boldsymbol{\psi})]\right) \\ &+ E_{q(\mathbf{W})q(\boldsymbol{\theta})}\left(\log[p(\mathcal{C}|\boldsymbol{\Omega}, \mathbf{W}, \boldsymbol{\psi}_D, \boldsymbol{\theta})]\right) \\ &- \mathrm{DKL}(q(\mathbf{W})\|p(\mathbf{W})) - \mathrm{DKL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})). \end{aligned} \tag{7}$$

The distribution $q(\mathbf{W})$ acts as a variational approximation and is assumed to be Gaussian, factorizing completely across weights and layers (l):

$$q(\mathbf{W}) = \prod_{j,k,l} p(W_{jk}^{(l)}) = \prod_{j,k,l} \mathcal{N}\left(m_{jk}^{(l)}, (s^2)_{jk}^{(l)}\right). \tag{8}$$

Extensions to approximations where we relax the factorization assumption are possible. Similarly, we are going to assume $q(\boldsymbol{\theta})$ to be Gaussian, and will assume no factorization, so that $q(\boldsymbol{\theta}) = \mathcal{N}(\mu_\theta, \Sigma_\theta)$.

4 Experiments

This section reports an in-depth validation of the proposed method on a variety of benchmarks. We are going to study the proposed variational framework for constrained dynamics in DGP models for ODE parameter estimates using equality constraints, and compare it against state-of-the-art methods. We will then consider the application of inequality constraints for a regression problem on counts, which was previously considered in the literature of monotonic GPs.

4.1 Settings for the proposed constrained dgp

We report here the configuration that we used across all benchmarks for the proposed method. Due to the generally low sample size n used across experiments (in most cases $n < 50$), unless specified otherwise the tests were performed with a two-layer DGP $\mathbf{f}(t) = \mathbf{f}^{(2)} \circ \mathbf{f}^{(1)}(t)$, with dimension of the "hidden" GP layer $\mathbf{f}^{(1)}(t)$ equal to 2, and RBF kernels. The length-scale of the RBF covariances was initialized to $\lambda_0 = \log(t_{\max} - t_{\min})$, while the marginal standard deviation to $\alpha_0 = \log(y_{\max} - y_{\min})$; the initial likelihood noise was set to $\sigma_0^2 = \alpha_0/10^5$. Finally, the initial ODE parameters were set to the value of 0.1. The optimization was carried out through stochastic gradient descent with Adaptive moment Estimation (Adam) Kingma & Ba (2017), through the alternate optimization of i) the approximated posterior over $\mathbf{W}$ and likelihood/covariance parameters ($q(\mathbf{W})$ and $\boldsymbol{\psi}$), and ii) likelihood parameters of ODE constraints and the approximate posterior over ODE parameters ($\boldsymbol{\psi}_D$ and $q(\boldsymbol{\theta})$). We note that the optimization of the ODE constraints parameters (the noise and scale parameters for Gaussian and Student-t likelihoods, respectively) is aimed at identifying in a fully data-driven manner the optimal trade-off between data attachment (likelihood term) and regularity (constraints on the dynamics). In what follows, DGP-t and DGP-G respectively denote the model tested with Student-t and Gaussian noise models on the ODE constraints.

4.2 Equality constraints from ODE systems

The proposed framework was tested on a set of ODE systems extensively studied in previous works: Lotka-Volterra (Goel et al., 1971), FitzHugh-Nagumo (FitzHugh, 1955), and protein biopathways from Vyshemirsky & Girolami (2007). For each experiment, we used the experimental setting

proposed in previous studies (Niu et al., 2016; Macdonald & Husmeier, 2015). In particular, for each test, we identified two experimental configurations with increasing modeling difficulty (e.g. less samples, lower signal-to-noise ratio, ...). A detailed description of the models and testing parameters is provided in the supplementary material. The experimental results are reported for parameter inference and model estimation performed on 5 different realizations of the noise.

4.2.1 Benchmark

We tested the proposed method against several reference approaches from the state-of-art to infer parameters of ODE systems.

RKG3: We tested the method presented in Niu et al. (2016) using the implementation in the `R` package `KGode`. This method implements gradient matching, where the interpolant is modeled using functions in Reproducing Kernel Hilbert spaces. This approach, for which ODE parameters are estimated and not inferred, was shown to achieve state-of-the-art performance on a variety of ODE estimation problems. We used values ranging from 10^{-4} to 1 for the parameter λ that the method optimizes using cross-validation.

Warp: In the `R` package `KGode` there is also an implementation of the warping approach presented in Niu et al. (2018). This method extends gradient matching techniques by attempting to construct a warping of the input where smooth Reproducing Kernel Hilbert spaces-based interpolants can effectively model nonstationary observations. The warping attempts to transform the original signal via assumptions on periodicity and regularity conditions. We used the default parameters and initialized the optimization of the warping function from a period equal to the interval where observations are available. Similarly to RKG3, ODE parameters are estimated and not inferred.

AGM: We report results on the Approximate Gradient Matching (AGM) approach in Dondelinger et al. (2013), implemented in the recently released `R` package `deGradInfer`. AGM implements a population Markov chain Monte Carlo approach tempering from the prior to the approximate posterior of ODE parameters based on an interpolation with GPs. In the experiments we use 10 parallel chains and we run them for 10^4 iterations. In the implementation of AGM, the variance of the noise on the observations is assumed known and it is fixed; we expect this to give a slight advantage to this method.

MCMC: In the `R` package `deGradInfer` there is also an implementation of a population Markov chain Monte Carlo sampler where the ODE is solved explicitly. In this case too we use 10 parallel chains that we run for 10^4 iterations. In contrast to AGM, in this implementation, the variance of the noise on the observations is learned together with ODE parameters.

4.2.2 Results

Figure 4.2.2 shows the distribution of the root mean squared error (RMSE) across folds for each experimental setting (see supplement for details). We note that the proposed method consistently leads to better RMSE values compared to the reference approaches (except some folds in one of the Fitz-Hugh-Nagumo experiments, according to a Mann-Whitney nonparametric test), and that DGP-t provides more consistent parameter estimates than DGP-G. This latter result may indicate a lower sensitivity to outliers derivatives involved in the functional constraint term. This is a crucial aspect due to the generally noisy derivative terms of nonparametric regression models. The distribution of the parameters for all the datasets tested in this study, which we report in the

supplementary material, reveals that, unlike the nonprobabilistic methods RKG3 and WARP, our approach is capable of inferring ODE parameters yielding meaningful uncertainty estimation.

4.3 Scalability test - large n

We tested the scalability of the proposed method with respect to sample size. To this end, we repeated the test on the Lotka-Volterra system with $n = 20, 40, 80, 150, 500, 10^3$, and 10^4 observations. For each instance of the model, the execution time was recorded and compared with the competing methods. All the experiments were performed on a 1.3GHz Intel Core i5 MacBook. The proposed method scales linearly with n (Figure 2), while it has an almost constant execution when $n < 500$; we attribute this effect to overheads in the framework we used to code our method. For small n, the running time of our method is comparable with competing methods, and it is considerably faster in case of large n.

4.4 Scalability test - large s

In order to assess the ability of the framework to scale to a large number of ODEs, we tested our method on the Lorenz96 system with increasing number of equations, $s = 125$ to $s = 1000$ (Lorenz & Emanuel, 1998). To the best of our knowledge, the solution of this challenging problem via gradient matching approaches has only been previously attempted in Gorbach et al. (2017). We could not find an implementation of their method to carry out a direct comparison, so we are going to refer to the results reported in their paper. The system consists of a set of drift states functions $(f_1(\mathbf{x}(t), \theta), f_2(\mathbf{x}(t), \theta), \ldots, f_s(\mathbf{x}(t), \theta))$ recursively linked by the relationship:

$$f_i(\mathbf{x}(t), \theta) = (x_{i+1}(t) - x_{i-2}(t))x_{i-1}(t) - x_i(t) + \theta,$$

where $\theta \in \mathbb{R}$ is the drift parameter. Consistently with the setting proposed in Gorbach et al. (2017); Vrettas et al. (2015), we set $\theta = 8$ and generated 32 equally spaced observations over the interval $[0, 4]$ seconds, with additive Gaussian noise $\sigma^2 = 1$. We performed two tests by training (i) on all the states, and (ii) by keeping one third of the states as unobserved, and by applying our method to identify model dynamics on both observed and unobserved states.

Figure 3 shows the average RMSE in the different experimental settings. As expected, the modeling accuracy is sensibly higher when trained on the full set of equations. Moreover, the RMSE is lower on observed states compared to unobserved ones. This is confirmed by visual inspection of the modeling results for sample training and testing states (Figure 4). The observed states are generally associated with lower uncertainty in the predictions and by an accurate fitting of the solutions (Figure 4, top). The model still provides remarkable modeling results on unobserved states (Figure 4, bottom), although with decreased accuracy and higher uncertainty. We are investigating the reasons for the posterior distribution over θ not covering the true value of the parameter across different experimental conditions.

4.4.1 Deep vs shallow

We explore here the capability of a DGP to accommodate for the data nonstationarity typical of ODE systems. In particular, the tests are performed in two different settings with large and small sample size n. By using the same experimental setting of Section 4.1, we sampled 80 and 1000 points,

respectively, from the FitzHugh-Nagumo equations. The data is modeled with DGPs composed by one ("shallow" GP), two and three layers, all with RBF covariances.

Table 1: Shallow and deep GP models under different experimental conditions in FitzHugh-Nagumo equations. Best results are highlighted in bold.

AVERAGE RMSE ACROSS PARAMETERS							
	RBF LAYERS			MATÉRN ν			
n	1	2	3	1/2	1	3/2	5/2
80	0.86	0.85	2.16	0.97	**0.72**	0.79	0.77
1000	0.66	**0.52**	0.53	0.87	0.63	0.70	0.74
DATA FIT RMSE							
80	0.23	**0.19**	0.42	0.23	0.20	0.23	0.25
1000	0.22	**0.17**	0.19	0.21	0.21	0.19	0.24

Figure 5 shows the modeling results obtained on the two configurations. We note that the shallow GP consistently underfits the complex dynamics producing smooth interpolants. On the contrary, DGPs provide a better representation of the nonstationarity. As expected, the three-layer DGP leads to sub-optimal results in the low-sample size setting. Furthermore, in order to motivate the importance of nonstationarity, which we implement through DGPs, we further compared against shallow GPs with lower degrees of smoothness through the use of Matérn covariances with degrees $\nu = 1/2, 1, 3/2, 5/2$.

The overall performance in parameter estimation and data fit is reported in Table 1. According to the results, a two-layer DGP provides the best solution overall in terms of modeling accuracy and complexity. Interestingly, the Matérn covariance, with an appropriate degree of smoothness, achieves superior performance in parameter estimation in case of low sample size. However, the nonstationarity implemented by the DGP outperforms the stationary Matérn in the data fit, as well as in the parameter estimation when the sample size is large. For an illustration of the data fit of the Matérn GP we refer the reader to the supplementary material. Crucially, these results indicate that our approach provides a practical and scalable way to learn nonstationarity within the framework of variational inference for deep probabilistic models.

4.5 Inequality constraints

We conclude our experimental validation by applying monotonic regression on counts as an illustration of the proposed framework for inequality constrains in DGP models dynamics. We applied our approach to the mortality dataset from Broffitt (1988), with a two-layer DGP initialized with an analogous setting to the one proposed in Section 4.1. In particular, the sample rates were modeled with with a Poisson likelihood of the form $p(y_i|\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$, and link function $\mu_i = \exp(f(t_i))$. Monotonicity on the solution was strictly enforced by setting $\psi_D = 5$. Figure 6 shows the regression results without (top) and with (bottom) monotonicity constraint. The effect of the constraint on the dynamics can be appreciated by looking at the distribution of the derivatives

(right panel). In the monotonic case the GP derivatives lie on the positive part of the plane. This experiment leads to results compatible with those obtained with the monotonic GP proposed in Riihimäki & Vehtari (2010), and implemented in the GPstuff toolbox Vanhatalo et al. (2013). However, our approach is characterized by appealing scalability properties and can implement monotonic constraints on DGPs, which offer a more general class of functions than GPs.

5 Conclusions

We introduced a novel generative formulation of deep probabilistic models implementing "soft" constraints on functions dynamics. The proposed approach was extensively tested in several experimental settings, leading to highly competitive results in challenging modeling applications, and favorably comparing with the state-of-the-art in terms of modeling accuracy and scalability. Furthermore, the proposed variational formulation allows for a meaningful uncertainty quantification of both model parameters and predictions. This is an important aspect intimately related to the application of our proposal in real scenarios, such as in biology and epidemiology, where data is often noisy and scarce.

Although in this study we essentially focused on the problem of ODE parameters inference and monotonic regression, the generality of our approach enables several other applications that will be subject of future investigations. We will focus on the extension to manifold valued data, such as spatio-temporal observations represented by graphs, meshes, and 3D volumes, occurring for example in medical imaging and system biology.

Acknowledgements

This work has been supported by the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01 (project Meta-ImaGen). MF gratefully acknowledges support from the AXA Research Fund.

This work is dedicated to Mattia Filippone.

References

Alvarez, M. A., Luengo, D., and Lawrence, N. D. Linear latent force models using Gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2693–2705, 2013.

Barber, D. and Wang, Y. Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1485–1493, Bejing, China, June 2014. PMLR.

Broffitt, J. D. Increasing and increasing convex Bayesian graduation. *Transactions of the Society of Actuaries*, 40(1):115–48, 1988.

Calandra, R., Peters, J., Rasmussen, C. E., and Deisenroth, M. P. Manifold Gaussian Processes for regression. In *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*, pp. 3338–3345, 2016.

Calderhead, B., Girolami, M., and Lawrence, N. D. Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 21*, pp. 217–224. Curran Associates, Inc., 2009.

Campbell, D. and Steele, R. J. Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, 22(2):429–443, March 2012.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. Random feature expansions for deep Gaussian processes. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 884–893, International Convention Centre, Sydney, Australia, August 2017. PMLR.

Da Veiga, S. and Marrel, A. Gaussian process modeling with inequality constraints. *Annales de la faculté des sciences de Toulouse Mathématiques*, 21(3):529–555, April 2012.

Damianou, A. C. and Lawrence, N. D. Deep Gaussian Processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Proceedings*, pp. 207–215. JMLR.org, 2013.

Dondelinger, F., Husmeier, D., Rogers, S., and Filippone, M. Ode parameter inference using adaptive gradient matching with gaussian processes. In Carvalho, C. M. and Ravikumar, P. (eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 216–228, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

Donohue, M. C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R. G., Raman, R., Gamst, A. C., Beckett, L. A., Jack, C. R., Weiner, M. W., Dartigues, J.-F., et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 10(5):S400–S410, 2014.

FitzHugh, R. Mathematical models of threshold phenomena in the nerve membrane. *The bulletin of mathematical biophysics*, 17(4):257–278, 1955.

Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553): 452–459, May 2015.

Goel, N. S., Maitra, S. C., and Montroll, E. W. On the volterra and other nonlinear models of interacting populations. *Reviews of modern physics*, 43(2):231, 1971.

González, J., Vujačić, I., and Wit, E. Reproducing kernel Hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45:26–32, 2014.

Gorbach, N. S., Bauer, S., and Buhmann, J. M. Scalable Variational Inference for Dynamical Systems. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4809–4818. Curran Associates, Inc., 2017.

Groeneboom, P. and Jongbloed, G. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2014.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980.

Konukoglu, E., Relan, J., Cilingir, U., Menze, B. H., Chinchapatnam, P., Jadidi, A., Cochet, H., Hocini, M., Delingette, H., Jaïs, P., et al. Efficient probabilistic model personalization integrating uncertainty on data and parameters: Application to eikonal-diffusion models in cardiac electrophysiology. *Progress in biophysics and molecular biology*, 107(1):134–146, 2011.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Liang, H. and Wu, H. Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008. PMID: 19956350.

Lorenz, E. N. and Emanuel, K. A. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, 55(3):399–414, 1998.

Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., and Ourselin, S. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, 2017.

Macdonald, B. and Husmeier, D. Gradient Matching Methods for Computational Inference in Mechanistic Models for Systems Biology: A Review and Comparative Analysis. *Frontiers in Bioengineering and Biotechnology*, 3:180, 2015.

Macdonald, B., Higham, C., and Husmeier, D. Controversy in mechanistic modelling with Gaussian processes. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1539–1547, Lille, France, July 2015. PMLR.

Mašić, A., Srinivasan, S., Billeter, J., Bonvin, D., and Villez, K. Shape constrained splines as transparent black-box models for bioprocess modeling. *Computers & Chemical Engineering*, 99: 96–105, 2017.

Meyer, M. C. Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, 2(3):1013–1033, 2008.

Minka, T. P. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pp. 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Neal, R. M. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics)*. Springer, 1 edition, August 1996.

Niu, M., Rogers, S., Filippone, M., and Husmeier, D. Fast Parameter Inference in Nonlinear Dynamical Systems using Iterative Gradient Matching. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1699–1707, New York, New York, USA, June 2016. PMLR.

Niu, M., Macdonald, B., Rogers, S., Filippone, M., and Husmeier, D. Statistical inference in mechanistic models: time warping for improved gradient matching. *Computational Statistics*, 33 (2):1091–1123, Jun 2018.

Rahimi, A. and Recht, B. Random Features for Large-Scale Kernel Machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. Curran Associates, Inc., 2008.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society Series B*, 69(5):741–796, 2007.

Rasmussen, C. E. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Riihimäki, J. and Vehtari, A. Gaussian processes with monotonicity information. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 645–652, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR.

Salzmann, M. and Urtasun, R. Implicitly Constrained Gaussian Process Regression for Monocular Non-Rigid Pose Estimation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 2065–2073. Curran Associates, Inc., 2010.

Schober, M., Duvenaud, D. K., and Hennig, P. Probabilistic ODE solvers with Runge-Kutta means. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 739–747. Curran Associates, Inc., 2014.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14(Apr): 1175–1179, 2013.

Varah, J. M. A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, 1982.

Vrettas, M. D., Opper, M., and Cornford, D. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. *Physical Review E*, 91(1):012148, 2015.

Vyshemirsky, V. and Girolami, M. A. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2007.

Wheeler, M. W., Dunson, D. B., Pandalai, S. P., Baker, B. A., and Herring, A. H. Mechanistic hierarchical Gaussian processes. *Journal of the American Statistical Association*, 109(507): 894–904, 2014.

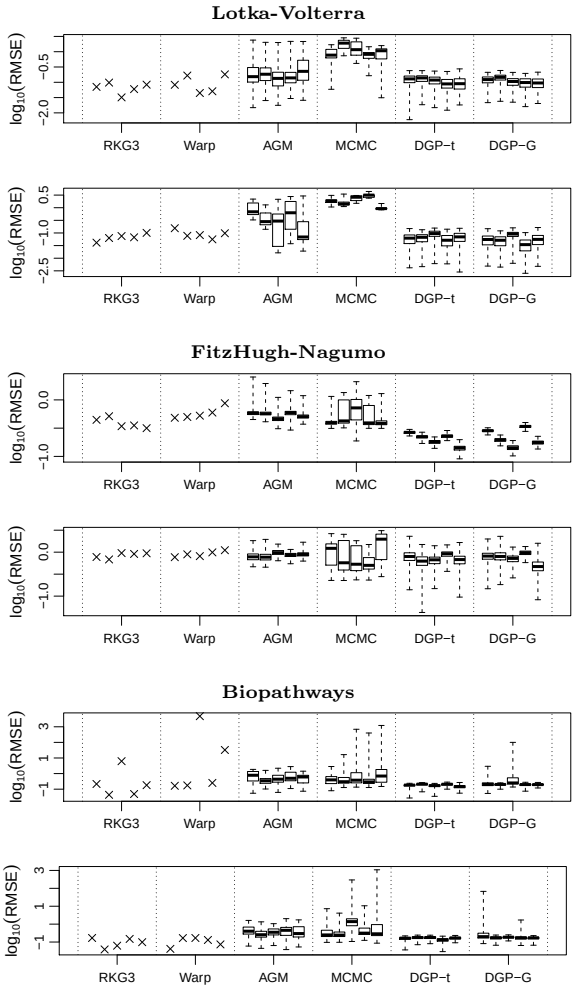


Figure 1: Boxplot of the RMSE on ODE parameters for the three ODE systems considered and for the two experimental settings. We report 5 bars for each method in the plots, corresponding to five different instantiations of the noise.

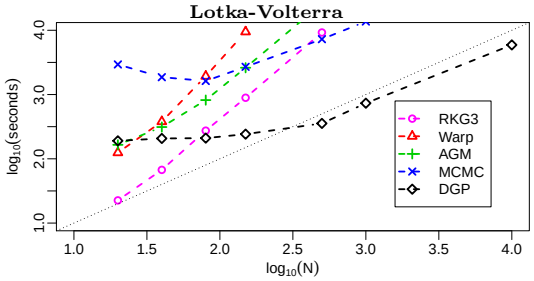


Figure 2: Execution time vs sample size – Lotka-Volterra ODE.

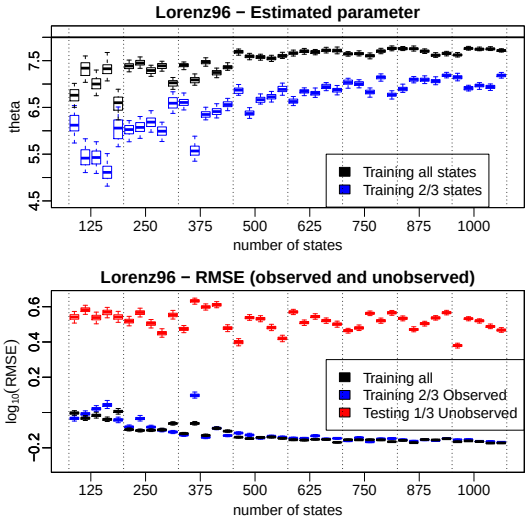


Figure 3: Top: parameter estimates in the different folds when training on all (black) or only on 2/3 (blue) of the states. The ground truth is indicated by the top horizontal bar ($\theta = 8$). Bottom: RMSE on the ODE curves fitting when training on all the states (black), and on the observed (blue) and unobserved (red) states when training on 2/3 of the states only.

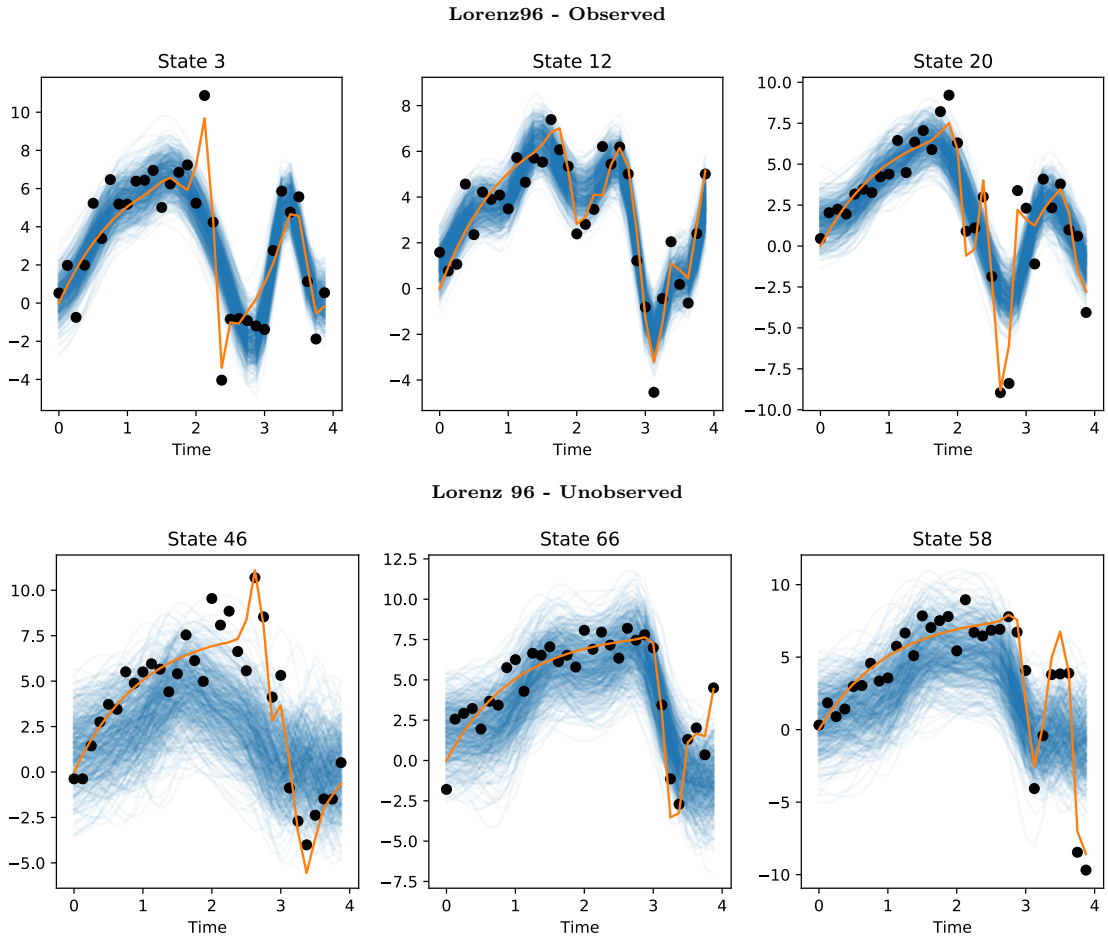


Figure 4: Model fit in Lorenz96. Randomly sampled observed (top) vs unobserved (bottom) states for $s = 125$ ODEs. Orange lines and black dots represent respectively the ground truth dynamics and noisy sample points. The blue lines are realizations of the DGP.

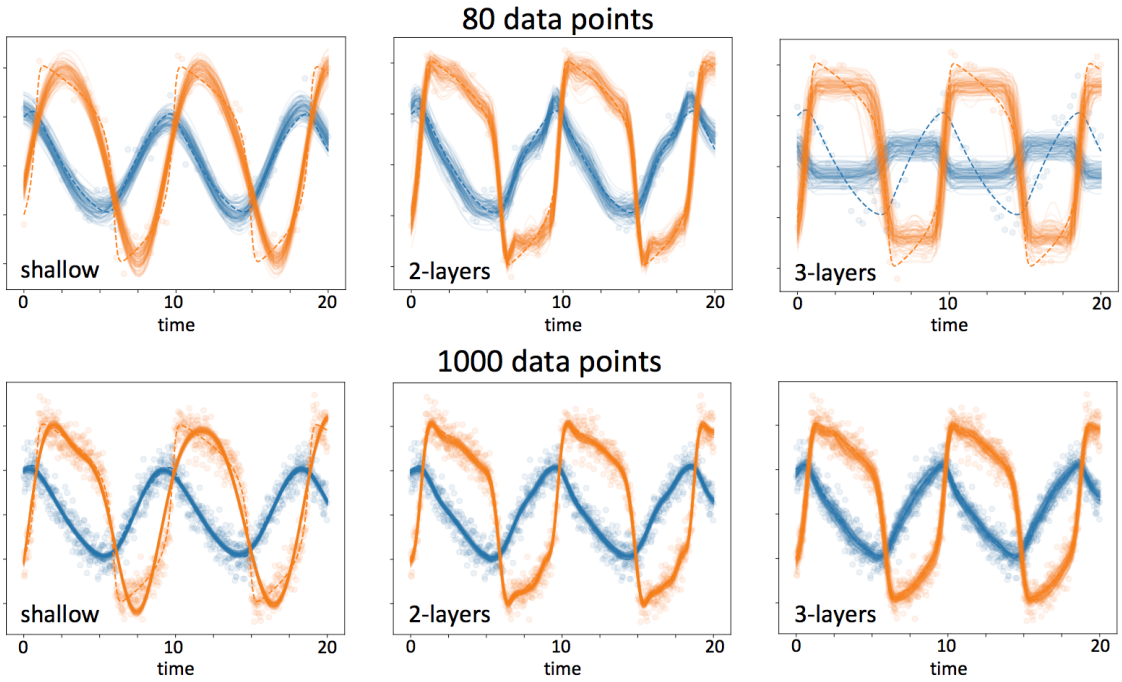


Figure 5: Modeling FitzHugh-Nagumo equations with GP and DGP. A deep model provides a more accurate description of data nonstationarity and associated dynamics (Table 1). Training points are denoted with circles; the ground truth trajectory is represented by the dashed line. Top: $N = 80$; Bottom: $N = 1000$. From left to right: Shallow GP, two-layers and three-layers DGP.

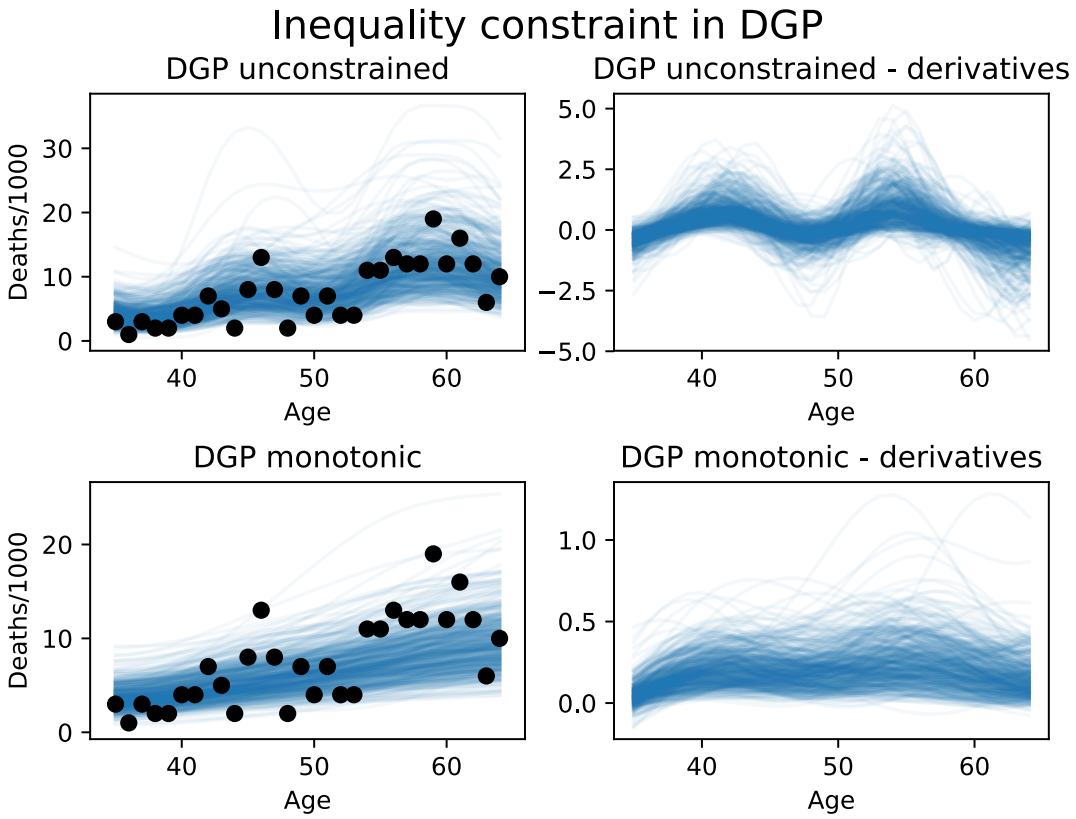


Figure 6: GP with Poisson likelihood: unconstrained (top) and monotonic (bottom). Black dots: observations from Broffitt (1988). Blue lines: GP realizations.

Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease.

Marco Lorenzi[1], Maurizio Filippone[2], Daniel C. Alexander[3] and Sebastien Ourselin[4]

1. EPIONE project-team, INRIA, Université Côte d'Azur, Sophia Antipolis, France
2. EURECOM, Department of Data Science, Sophia Antipolis, France
3. POND group, Centre for Medical Image Computing, University College London, UK
4. Translational Imaging Group, Centre for Medical Image Computing, University College London, UK

Originally published in:
NeuroImage, S1053-8119(17)30706-1, 2017

Abstract

BACKGROUND: We evaluate the use of probabilistic disease progression model of the natural history of Alzheimer's disease (AD) to automatically quantify the individual disease severity in the clinical scenario with respect to the modelled pathological progression.
METHODS: The AD progression was estimated on an ADNI cohort of 200 amyloid positive individuals, and the quality and uncertainty of the model predictions was assessed on an independent cohort of 582 individuals with respect to missing measurements, biomarkers, and follow-up information.
RESULTS: The estimated AD progression spans across roughly 20 years. The automatic staging of the model on testing individuals shows high face validity with respect to the clinical diagnosis (AUC ¿.87 for estimating conversion to mild cognitive impairment and to AD). The biomarkers allowing the most accurate identification of the disease staging were the clinical scores and hippocampal volume.
CONCLUSIONS: The disease progression model provides a statistical reference for assessing the pathological stage of de-novo individuals in the clinical setting, and represents a valuable instrument for quantifying the variability and the clinical value of the biomarkers across disease stages.

1 Introduction

Neurodegenerative disorders (NDDs), such as Alzheimer's disease (AD), are characterised by the progressive pathological alteration of the brain's biochemical processes and morphology, and ultimately lead to the irreversible impairment of cognitive functions [3]. The correct understanding of the relationship between the different pathological features is of paramount importance for improving the identification of pathological changes in patients, and for better treatment [12].

To this end, ongoing research efforts aim at developing precise models allowing optimal sets of measurements (and combinations of them) to uniquely identify pathological traits in patients. This problem requires the definition of optimal ways to integrate and jointly analyze the heterogeneous multi-modal information available to clinicians [37, 24, 20]. By consistently analyzing multiple biomarkers that to date have mostly been considered separately, we aim at providing a richer description of the pathological mechanisms and a better understanding of individual disease progressions.

Disease progression modeling is a relatively new research direction for the study of NDD data [6, 14, 5, 35, 2, 30, 9, 22]. The main goal of disease progression modeling consists in revealing the natural history of a disorder from collections of imaging and clinical data by: 1) *quantifying* the dynamics of NDDs along with the related temporal relationship between different biomarkers, and 2) *staging* patients based on individual observations for diagnostic and interventional purposes. Therefore, this research domain is closely related to the exploitation of advanced statistical/machine-learning approaches for the joint modelling of the heterogeneous and information available to clinicians: imaging, biochemical, and clinical biomarkers. Differently from the several predictive machine-learning approaches proposed in the past in NDD research, disease progression models aim at explicitly estimating the temporal progression of the biomarker to provide a better interpretation and understanding of the natural evolution of the pathology. For this reason it represents a very appealing modeling approach in clinical settings.

The main challenge addressed by disease progression modelling consists in the general lack a well-defined temporal reference in longitudinal clinical datasets of NDDs. Indeed, age or visit date information are biased time references for the individual longitudinal measurements, since the onset of the pathology may vary across individuals according to genetic and environmental factors [34]. This is a very specific methodological issue requiring the extension and generalization of the analysis approaches classically used in time-series analysis.

To tackle this problem, it is usually assumed that individual biomarkers are measured relatively to an underlying disease trajectory defined with respect to an absolute time axis describing the natural history of the pathology [14]. Each individual is thus characterized by a specific observation time that needs to be estimated in order to assess the individual pathological stage. According to this statistical setting, we therefore aim at estimating a *group-wise* disease model defined with respect to an absolute time scale, along with *individual* time re-parameterization relative to the group-wise progression.

This modeling paradigm has been implemented in a number of approaches proposed in the recent years. In [5] the authors proposed to model the temporal biomarker trajectories through random effect regression, building on the established theory of self-modeling regression [16], while the authors of [30] re-frame the random effect regression model in a geometrical setting, based on the assumption of a logistic curve shape for the average biomarker trajectories.

Finally, progression models have been recently extended to the modelling of brain images based on the time-reparameterization of voxel/mesh-based measures [35, 2, 22].

While the focus of these works is essentially on the estimation of average progression trajectories from individual time-series, they generally lack a probabilistic formulation, which makes it difficult to quantify the uncertainty associated with the biomarkers trajectories, as well as to predict the pathological stage of *de-novo patients*. This is a very crucial aspect for the use of disease progression models in the clinical setting, for disease severity quantification and diagnostic purposes.

The use of disease progression models for predictive purposes is indeed much less investigated, for example for identifying the clinical progression of an individual. Predictive models of patient staging were proposed within the setting of the Event Based Models [6], or still through random effect modeling [9]. However, the event based model relies on the coarse binary discretization of the biomarker changes, and does not account for longitudinal observations, while the predictive models proposed in [9] and [31] require cohorts with known disease onset, and therefore lacks flexibility while being prone to bias due to mis-diagnosis and uncertainty of the conversion time.

Nevertheless, the ensemble of this research offers a sight of the potential of these approaches in representing a novel and powerful diagnostic instrument in the clinical scenario: in this study we thus aim at assessing the clinical use of disease progression modeling in Alzheimer's disease. To this end, we reformulate disease progression modeling within a Bayesian setting in order to allow the probabilistic estimate of the biomarker trajectories and the quantification of the uncertainty of predictions of the individual pathological stage.

Thanks to these rich statistical properties we quantify the ability of the model of disease progression in providing accurate estimates of the pathological stage in de-novo individuals, while accounting for potential missing observations, biomarkers, or follow-up information. Moreover, we show that disease progression modeling allows quantifying the clinical value of the biomarkers across different clinical settings and pathological stages.

The manuscript is structured as follows. Section 2.1 formulates disease progression modeling based on Bayesian Gaussian Process regression [27], while Section 2.2 illustrates the validation of our model on clinical and multivariate imaging measurements from a cohort of 782 amyloid positive individuals extracted from the ADNI database.

2 Methods

2.1 Statistical setting

This section highlights the statistical framework employed in this study, based on the reformulation of self-modeling regression withing a Bayesian setting. This achieved by 1) defining a random effect Gaussian process regression model to account for individual correlated time series (section 2.1); 2) modeling individual time transformations encoding the information on the latent pathological stage (section 2.1.2); and 3) introducing a monotonicity information in order to impose a regular behaviour on the biomarkers trajectories (section 2.1.3). We finally illustrate in section 2.1.4 how the proposed framework leads to a probabilistic model of disease staging in de-novo individuals, naturally accounting for missing information. Further details on model specification and inference are provided in the Supplementary Section 6.1, while the experimental validation on synthetic data is reported in Supplementary Section 6.2.

2.1.1 Gaussian process-based random effect modeling of longitudinal progressions

In what follows, longitudinal measurements of N_b biomarkers $\{b_1, \dots, b_{N_b}\}$ over time are given for N individuals.

We represent the longitudinal biomarker's measures associated with each individual j as a multidimensional array $(\mathbf{y}^j(t_1), \mathbf{y}^j(t_2), \dots, \mathbf{y}^j(t_{k^j}))^\top$ sampled at k^j multiple time points $\mathbf{t} = \{t_1, t_2, \dots, t_{k^j}\}$. Although different biomarkers may be in reality sampled at different time-points, for the sake of notation simplicity in what follows we will assume, without loss of generality, that the sampling time is common among them. The observations for individual j at a single time point t are thus a random sample from the following generative model:

$$\mathbf{y}^j(t) = \left(y^j_{b_1}(t), y^j_{b_2}(t), \dots, y^j_{b_{N_b}}(t)\right)^\top \tag{1}$$

$$= \mathbf{f}(t) + \boldsymbol{\nu}^j(t) + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{f}(t) = (f_{b_1}(t), f_{b_2}(t) \dots, f_{b_{N_b}}(t))^\top$ is the fixed effect function modelling the biomarker's longitudinal evolution, $\boldsymbol{\nu}^j(t) = (\nu^j_{b_1}(t), \nu^j_{b_2}(t), \dots, \nu^j_{b_{N_b}}(t))^\top$ is the individual random effect, and $\boldsymbol{\epsilon} = (\epsilon_{b_1}, \epsilon_{b_2}, \dots, \epsilon_{b_{N_b}})^\top$ is time-independent observational noise. The group-wise evolution is modelled as a zero-mean GP, $\mathbf{f} \sim \mathcal{GP}(0, \Sigma_G)$, the individual random effects are assumed to be Gaussian distributed correlated signals $\boldsymbol{\nu}^j \sim \mathcal{N}(0, \Sigma_S)$, while the observational noise is assumed to be a Gaussian heteroskedastic term $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma_\epsilon)$, where Σ_ϵ is a diagonal matrix $\text{diag}[\boldsymbol{\sigma}^2_{b1}, \boldsymbol{\sigma}^2_{b2}, \dots, \boldsymbol{\sigma}^2_{b_{N_b}}]$.

Fixed Effect Process

The covariance function Σ_G describes the biomarkers temporal variability, and is represented as a block-diagonal matrix
$\Sigma_G(\mathbf{f}, \mathbf{f}) = \text{diag}[\Sigma_{b_1}(\mathbf{f}_{b_1}, \mathbf{f}_{b_1}), \Sigma_{b_2}(\mathbf{f}_{b_2}, \mathbf{f}_{b_2}), \dots, \Sigma_{b_{N_b}}(\mathbf{f}_{b_{N_b}}, \mathbf{f}_{b_{N_b}})]$, where each block represents the within-biomarker temporal covariance expressed as a negative squared exponential function $\Sigma_b(\mathbf{f}_b(t_1), \mathbf{f}_b(t_2)) = \eta_b \exp\left(-\frac{(t_1-t_2)^2}{2\, l_b^2}\right)$, and where the parameters η_b and l_b are the marginal variance and length-scale of the biomarker's temporal evolution, respectively.

Individual Random Effects

The random covariance function Σ_S models the individual deviation from the fixed effect, and is represented as a block-diagonal matrix $\Sigma_S(\boldsymbol{\nu}^j, \boldsymbol{\nu}^j) = \text{diag}[\,\Sigma^j_{b_1}(\boldsymbol{\nu}^j_{b_1}, \boldsymbol{\nu}^j_{b_1}), \Sigma^j_{b_2}(\boldsymbol{\nu}^j_{b_2}, \boldsymbol{\nu}^j_{b_2}), \dots, \Sigma_{b^j_{N_b}}(\boldsymbol{\nu}^j_{b_{N_b}}, \boldsymbol{\nu}^j_{b_{N_b}})]$, where each block Σ^j_b corresponds to the covariance function associated with the individual process $\boldsymbol{\nu}^j_b(t)$. Thanks to the flexibility of the proposed generative model, any form of the random effect covariance Σ_S can be easily specified in order to model the subject-specific biomarkers' progression. In what follows we will use a linear covariance form $\Sigma^j_b(\boldsymbol{\nu}^j_b(t_1), \boldsymbol{\nu}^j_b(t_2)) = (\sigma^j_b)^2\left((t_1 - \bar{\mathbf{t}})(t_2 - \bar{\mathbf{t}})\right)$, where $\bar{\mathbf{t}}$ is the average observational time for individual j, when more than 4 measurements are available, and i.i.d. Gaussian covariance form $\Sigma^j_b(\boldsymbol{\nu}^j_b(t_1), \boldsymbol{\nu}^j_b(t_2)) = (\sigma^j_b)^2$ when 2 or 3 measurements are available, while assigning it to 0 otherwise. This choice is motivated by stability concerns, in order to keep the model complexity compatible with the generally limited number of measurements available for each individual.

2.1.2 Individual time transformation

The generative model (1) is based on the key assumption that the longitudinal observations across different individuals are defined with respect to the same temporal reference. This assumption may be invalid when the temporal alignment of the individual observations with respect to the common group-wise model is unknown, for instance in the typical scenario of a clinical trial in AD where the patients' observational time is relative to the common baseline, and where the disease onset is a latent event (past or future) which is not directly measurable. We assume that each individual measurement is made with respect to an absolute time-frame τ through a time-warping function $t = \phi^j(\tau)$ that models the time-reparameterization with respect to the common group-wise evolution. Model (1) can thus be reparameterized as

$$\mathbf{y}^j(\phi^j(\tau)) = \mathbf{f}(\phi^j(\tau)) + \boldsymbol{\nu}^j(\phi^j(\tau)) + \boldsymbol{\epsilon}. \tag{3}$$

The present formulation allows the specification of any kind of time transformation, and in what follows we shall focus on the modelling of a linear reparameterization of the observational time $\phi^j(\tau) = \tau + d^j$. This modeling assumption is mostly motivated by the choice of working with a reasonably limited number of parameters, compatibly with the generally short follow-up time available per individual (cfr. Table 2). Within this setting, the time-shift d^j encodes the disease stage associated with the individual relatively to the group-wise model.

Model parameters

Overall, model (3) is identified by $(N_j + 3)N_b + N_j$ parameters, represented by the fixed effects and noise $\boldsymbol{\theta}_G = \{\eta_{b_k}, l_{b_k}, \epsilon_{b_k}\}_{k=1}^{N_b}$, by the individual random effects parameters $\boldsymbol{\theta}_G^j = \{\sigma_{b_k}^j\}_{k=1}^{N_b}$ and by the time-shifts d^j.

In what follows, the optimal parameters are obtained by maximising $\log \mathcal{L}$ through conjugate gradient descent, via alternate optimization between the hyper-parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$, and the individuals' time-shifts d^j. Regularization was also enforced by introducing Gaussian priors for the parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$.

2.1.3 Monotonic constraint in random-effect multimodal GP regression

Due to the non-parametric nature of Gaussian process regression, we need an additional constraint on model (3) in order to identify a unique solution for the time transformation. By assuming a steady temporal evolution of biomarkers from normal to pathological values, we shall assume that the biomarker trajectories described by (3) follow a (quasi) monotonic behaviour. This requirement can be implemented by imposing a prior positivity constraint on the derivatives of the GP function. Inspired by [28], we impose a monotonicity constraint by assuming a probit-likelihood for the derivative measurements $\mathbf{m}(t)$ associated with the derivative process $\dot{\mathbf{f}}(t) = \frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t}$ at time t:

$$p(\mathbf{m}(t)|\dot{\mathbf{f}}(t)) = \Phi\left(\frac{1}{\lambda}\dot{\mathbf{f}}(t)\right), \tag{4}$$

with $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x|0,1)\,dx$. The quantity $\lambda > 0$ is an additional model parameter controlling the degree of positivity enforced on the derivative process, with values approaching zero for stronger

monotonicity constraint. In what follows, the monotonicity of each biomarker is controlled by placing 10 derivative points equally spaced on the observation domain, and by fixing the N_b derivative parameters $\{\lambda_{b_k}\}_{k=1}^{N_b}$ to the value of 1e-6. The position of the derivative points was updated at each iteration, according to the changes of the GP domain.

By following a similar construction, we could equally enforce a monotonic behavior to the random effects associated with the individual trajectories. This additional constraint would however come with a cumbersome increase of the model complexity, since it would introduce an additional layer of virtual derivative parameters (with associated location) per individual. Moreover, while we are interested in modeling a globally monotonic biomarker trajectory on the fixed parameters, we relax this constraint at the individual level, since some subjects may be characterised by non strictly monotonic time-series due to specific clinical conditions.

2.1.4 Prediction of observations and individual staging

Gaussian processes naturally allow for probabilistic predictions given the observed data. At any given time point t^*, the posterior biomarker distribution has the Gaussian form $p(\mathbf{f}^*|t^*, \mathbf{y}, t, \mathbf{m}, t') \sim \mathcal{N}(\mathbf{f}^*|\boldsymbol{\mu}^*, \Sigma^*)$:

$$\begin{aligned} \boldsymbol{\mu^*} &= \Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t))(\Sigma_{\text{joint}} + \tilde{\Sigma}_{\text{joint}})^{-1}\tilde{\boldsymbol{\mu}}_{\text{joint}} \quad (5) \\ \Sigma^* &= \Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t^*)) - \Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t)) \\ &\quad (\Sigma_{\text{joint}} + \tilde{\Sigma}_{\text{joint}})^{-1}\Sigma_G(\mathbf{f}(t), \mathbf{f}(t^*)), \quad (6) \end{aligned}$$

where the matrix $\left(\Sigma_{\text{joint}} + \tilde{\Sigma}_{\text{joint}}\right)$ is the joint covariance resulting from the inference scheme detailed in Supplmentary Section 6.1 [28].

We also derive a probabilistic model for the individual temporal staging given a set of biomarker observations $\mathbf{y}^*$, thanks to the Bayes formula:

$$\begin{aligned} p(t^*|\mathbf{y}^*, \mathbf{y}, t, \mathbf{m}, t') &= p(\mathbf{y}^*|t^*, \mathbf{y}, t, \mathbf{m}, t') \\ &\quad p(t^*)/p(\mathbf{y}^*|\mathbf{y}, t, \mathbf{m}, t'), \quad (7) \end{aligned}$$

which we compute by assuming an uniform distribution on t^*, and by noting that $p(\mathbf{y}^*|t^*, \mathbf{y}, t, \mathbf{m}, t') \sim \mathcal{N}(\boldsymbol{\mu}^*, \Sigma^* + \Sigma_\epsilon)$. In particular, the covariance form $\Sigma_G(\mathbf{f}(t^*), \mathbf{f}(t^*))$ can be specified in order to account for incomplete data, and thus generalizes the GP model for predictions in presence of *missing biomarker observations*. The posterior distribution (7) quantifies the *confidence* of the model about the individual disease staging, and thus is a valuable information about the precision of the diagnosis. We will also compute the *expectation* of the distribution $p(t^*|\mathbf{y}^*, \mathbf{y}, t, \mathbf{m}, t')$, which provides a scalar value that can be used in subsequent classification methods.

2.2 Participants and Study Data

Data used in the preparation of this article were obtained from the ADNI database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic

Group	NL	NL converted	MCI stable	MCI converted	AD
Training data					
N	67	5	0	53	75
Age	73 (6)	81.4 (5.2)	/	72 (7.7)	73 (8.5)
Sex (% females)	61	0	/	43	45
Education (yrs)	16.2 (2)	17.2 (3)	/	15.8 (2.6)	16 (2.4)
ADAS13	8.8 (4.5)	13.8 (2.4)	/	22.6 (6.7)	31.3 (8.5)
FAQ	0.2 (0.6)	0.4 (0.5)	/	5.2 (4.5)	12.9 (7)
RAVLT learning	5.6 (2.6)	2.2 (1.9)	/	3.2 (2.5)	1.8 (1.7)
Entorhinal (cm^3)	3.9 (0.6)	3.7 (0.5)	/	3.2 (0.7)	2.9 (0.6)
Hippocampus (cm^3)	7.5 (0.9)	6.7 (0.7)	/	6.2 (0.9)	6 (9.3)
Ventricles (cm^3)	36 (20)	57 (26)	/	42 (21)	47 (22)
Whole brain (cm^3)	1057 (105)	1106 (116)	/	1040 (107)	1013 (113)
FDG	6.6 (0.5)	6.1 (0.65)	/	5.7 (0.6)	5.2 (0.64)
AV45	1.2 (0.2)	1.3 (0.09)	/	1.4 (0.2)	1.4 (0.2)
Testing data					
N	74	17	243	106	145
Age	75.3 (5.9)	76.5 (4)	73.3 (7)	73.6 (7.3)	75 (7.9)
Sex (% females)	55	41	39	40	39
Education (yrs)	16 (2.9)	16.2 (2.6)	16 (2.8)	16 (3)	15.3 (3.1)
ADAS13	9.8 (4)	11.7 (3.4)	15.7 (6)	21 (6.1)	29.4 (8.2)
FAQ	0.5 (1.3)	0.6 (1.6)	2.7 (3.5)	5.1 (4.7)	12.9 (6.8)
RAVLT learning	5.6 (2.2)	5.6 (2.7)	4.3 (2.5)	2.8 (2.2)	1.8 (1.9)
Entorhinal (cm^3)	3.8 (0.4)	3.6 (0.7)	3.6 (0.7)	3.1 (0.7)	2.7 (0.7)
Hippocampus (cm^3)	7.2 (0.7)	7.2 (0.8)	6.9 (1)	6 (0.8)	5.7 (0.1)
Ventricles (cm^3)	33 (15)	44 (21)	39 (23)	41 (23)	49 (24)
Whole brain (cm^3)	1019 (102)	1055 (93)	1056 (100)	992 (110)	972 (124)
FDG	6.5 (0.62)	6.4 (0.7)	6.3 (0.7)	5.9 (0.6)	5.4 (0.7)
AV45	1.21 (0.19)	1.4 (0.2)	1.3 (0.19)	1.4 (0.2)	1.4 (0.2)

Table 1: Baseline sociodemographic and clinical information for training and testing study cohort. NL: normal individuals, MCI: mild cognitive impairment, AD: Alzheimer's patients.

resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

We collected longitudinal measurements for the ADNI individuals with baseline values of CSF Aβ amyloid lower than the nominal values of 192 pg/ml. The information was extracted from the ADNIMERGE[1] R package[26] (MEDIAN field of the UPENNBIOMK_MASTER table). This preliminary selection is aimed to validate the model on a clinical population likely to represent the whole disease time-span.

[1] `adni.bitbucket.io/adnimerge.html`

2.2.1 Data processing

The model was trained on a group of 200 randomly selected individuals including healthy volunteers, mild cognitive impairment subjects converted to AD (MCI conv), and AD patients having at least one measurement for each of the following biomarkers: *volumetric measures* (hippocampal, ventricular, entorhinal, and whole brain volumes), *glucose metabolism* (average normalized FDG uptake in prefrontal cortex, anterior cingulate, precuneus and parietal cortex), *brain amyloidosys* (average normalized AV45 uptake in frontal cortex, anterior cingulate, precuneus and parietal cortex), and *cognitive function* measured by common cognitive questionnaires (ADAS13, RAVLT learning score, FAQ). The testing set was composed by the remaining 582 subjects with at least a missing biomarker, as well as by the subgroup of MCI non converted to AD during the observational time (MCI stable). The volumetric measures were scaled by the individual total intracranial volume, and all the biomarkers measurements were converted into quantile scores (0 to 1 for normal to abnormal values), with respect to the biomarkers distribution of the *training set*. This latter modeling precaution is aimed to avoid spurious correlation between training and testing data due to the combined normalization of the values. Table 1 shows baseline clinical and sociodemographic information of the individuals used respectively in training and testing set, while in Table 2 we report the average follow-up time and the ratio of missing data of the pooled sample. Supplementary Section 6.2.4 reportes the R code used for the data pre-processing.

2.3 Longitudinal modelling of Alzheimer's disease progression

Model training

The model was applied in order to estimate the temporal biomarker evolution and the disease stage associated with each individual in the training set. The plausibility of the model was assessed by group-wise comparison of the predicted time-shift, and by correlation with respect to the time to AD diagnosis for the MCI individuals subsequently converted to AD. For sake of comparison we also compared the progression modelled with our approach with respect to the one estimated with the method proposed in [5]. The method was applied to the training data by using the standard parameters defined in the R package GRACE[2] (see Supplementary Material 6.2.2 for further details).

Clinical use of the progression model on de-novo individuals

The estimated disease progression model provides a valuable reference in the clinical setting, as it can be used to predict an individual pathological stage, as well as to quantify the biomarker predictive value, or the influence of missing data. To this end, we estimated the predictive performance of

[2]https://mdonohue.bitbucket.io/grace/

Ventr	Hippo	Ent	Whole Brain	ADAS13	FAQ	RAVLT	AV45	FDG
Training data								
2.3 (0)	2.3 (0)	2.3 (0)	2.3 (0)	3 (0)	3.3 (0)	3.3 (0)	1.9 (0)	1.6 (0)
Testing data								
3.4 (11)	3.4 (11)	3.4 (11)	3.4 (11)	3.9 (0)	3.9 (0)	3.9 (0)	3.8 (43)	3 (19)

Table 2: Average follow-up years and percentage of individuals with missing data (in parenthesis).

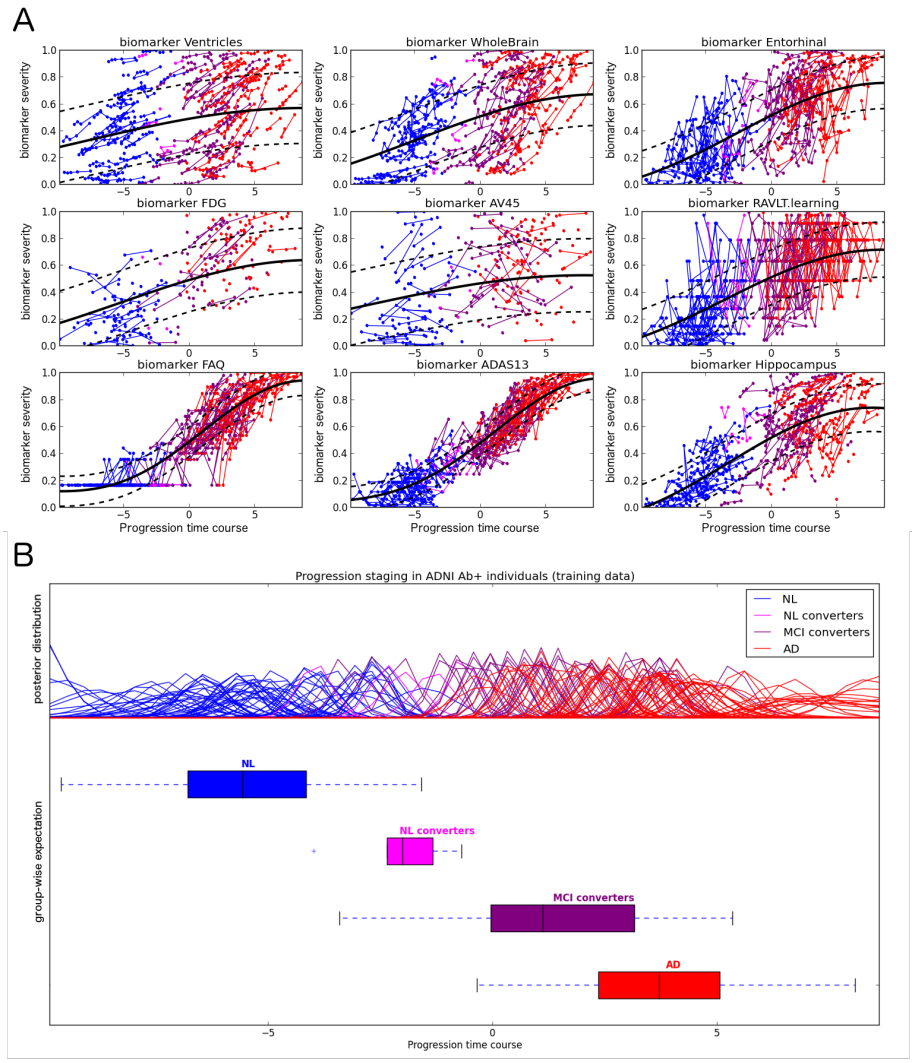


Figure 1: A) Modelled biomarker progression in the training set of 200 Aβ amyloid positive individuals (solid/dashed lines: mean $\pm$ sd). B) Posterior prediction for the individual time shift in training data. Healthy individuals are generally displaced at the early stages of the pathology, while the predictions for MCI and AD patients are associated with respectively intermediate and late progression stages. NL: normal individuals, MCI: mild cognitive impairment, AD: Alzheimer's patients.

the model in assessing the individual pathological stage with respect to follow-up assessments and missing biomarkers. This was done by estimating the predictive accuracy of the group-wise separation obtained via increasing thresholds of the estimated temporal progression.

2.4 Results

Model plausibility

The estimated biomarker progression (Figure 1-A) shows a biologically plausible description of the pathological evolution, compatible with previous findings in longitudinal studies in familial AD [1], and with the hypothetical models of AD progression [12, 7]. The progression is defined on a time scale spanning roughly 20 years, and is characterized at the initial stages by high-levels of AV45, followed by an increase in ventricles volume and abnormality of FDG uptake. These latter measures are however heterogeneously distributed across clinical groups, and with rather large variability. The evolution is further characterized by increasing abnormality of the volumetric measures, and by the steady worsening of neuropsychological scores such as FAQ.

Figure 1-B shows the posterior distribution associated with the modelled progression for each individual time shift. Healthy individuals (blue) are associated with the early stages of the pathology in both training and testing data, while MCI (purple) and AD patients (red) are characterized by respectively intermediate and late predicted progression stages. The group-wise comparison between the expected time-shifts was statistically significant between each group pairs (ANOVA, p <1e-6). Moreover, the time to conversion to AD in the MCI group was significantly correlated with the disease staging quantified by the expectation of the individual time distribution ($R^2 = -0.4$, $p = 3.8e-4$).

Finally, when applying [5] to the training data we measured a strong agreement between the resulting progression and the one obtained with our method, resulting in a correlation between the corresponding individual time-shifts of 0.94 (p <1e-6) (Supplementary Material 6.2.2).

Clinical use of the progression model

Figure 2 shows the individual posterior predictive distribution associated to the testing individuals (staging probability), and the boxplot of the group-wise distribution of the expected time-shift when using the model as statistical reference through formula (7). The figure reports the two different modeling scenarios based on baseline information only (top), and on the complete set of longitudinal measurements (bottom). Although the group-wise differences between the expected time-shifts are similar for both scenarios, the figure shows that the use of follow-up information largely reduces the uncertainty of the prediction. Indeed, the individual time distributions predicted by using multiple measurements are narrower as compared to the less informative confidence margins associated to the baseline information only.

As with the training case, for both scenarios the group-wise distribution of the expected time-shift shows a significant separation between the clinical groups according to the increase of the pathological stage (ANOVA, p <1e-4). Interestingly, the temporal positioning of the non converting MCI lies between controls and MCI converters, and is on average lower than the one of healthy individuals subsequently converted to cognitive impairment.

Figure 3 reports the the classification results based on the baseline information only, and on increasing thresholds of the progression time course. Although the model is not optimized to explicitly classify the clinical groups, the simple thresholding based on the model prediction generally shows high face validity with respect to the clinical diagnosis. For all the considered scenarios, the highest accuracy is reached in a time window around the point $t = 0$, while the area under the ROC curve is .99, .88 and .87 for NL vs AD, MCI converters vs MCI stable, and NL converters vs NL stable, respectively.

We further tested the model in presence of missing information, by computing the predictions when only one baseline biomarker is available (Figure 4). The predictive outcomes show important variations depending on the considered biomarker, while the confidence bounds for the predictions are usually large, to denote high uncertainty. Nevertheless, the group-wise boxplots of the expected time-shift show that FAQ, ADAS13, and hippocampal volume are the biomarkers leading to the largest group-wise separation. This aspect is quantified in Table 3, reporting the discrimination results with respect to the nominal cut-off point of $t = 1.65$, corresponding to the 15th percentile of the distribution of the expected time-shift in the training AD group. Although the highest discriminative results are obtained when the biomarkers are used jointly, the clinical tests generally lead to the best predictive performance. This is related to the lower uncertainty of the modelled progressions (Figure 2-A), which leads to a more accurate identification of the individual staging along the pathological trajectory.

These figures were similar when considering the single biomarkers within the longitudinal setting, where the clinical tests still outperformed the other biomarkers in discriminating the clinical groups (Supplementary Figure 6).

3 Discussion

Whilst most of current approaches to disease progression modeling mainly focused on describing the multivariate dynamics of the biomarkers trajectories, in this study we thoroughly assessed the use of disease progression modeling for predictive purposes and uncertainty quantification. In particular, we showed that the proposed model provides a statistical reference for assessing the pathological stage of testing individuals by optimally combining the information provided by the several biomarkers into a biologically plausible and intelligible score quantified by the time shift. The disease progression model itself thus can be seen as a novel biomarker of pathological progression.

We illustrated the use of the model as benchmarking tool for the statistical comparison of biomarkers. The model allows the quantification of the variability associated to the single biomarkers, and can be thus used as a reference for screening and enrichment purposes [19, 38, 11].

The modelled progression showed that clinical tests generally lead to lower uncertainty for identifying the individual clinical stage. This finding is compatible with the results reported by previous disease progression models applied to ADNI, such as [14] and [36]. In this latter study ADAS13 consistently appeared among the first events distinguishing the normal disease stages from the pathological ones. However, some care should be taken in drawing conclusions from the present analysis. For example, our model was based on the standard volumetric measures provided in the ADNI database, and we cannot exclude that the more precise quantification of morphological brain changes would lead to better performance of volumetric biomarkers [32, 4].

Thanks to the probabilistic formulation we showed that the use of longitudinal information is important for reducing the uncertainty of the prediction, and thus allows one to better identify the disease status associated to an individual. This important aspect is in agreement with the generally higher statistical precision reported in previous Alzheimer's studies comparing longitudinal measurements to baselines ones [10, 7, 33].

In this work we focused on the modelling of the progression of amyloid positive individuals. This choice was motivated by the interest in assessing the model performance on an homogeneous clinical population likely to be representative of the Alzheimer's evolution. While the absence of

pathological amylod levels seems indicative of non-Alzheimer's pathophysiology [8, 23], there is currently an active debate on the mechanisms of neurodegeneration not related to brain amyloidosis [13]. The investigation of these aspects goes beyond the scope of the present work, and future extensions of disease progression modeling will aim at identifying differential progressions underlying sub-pathologies, for example by reformulating the proposed random effect regression within the realm of Gaussian process mixture models [17, 29].

3.1 Methodological considerations

From the methodological perspective, we proposed a novel probabilistic approach based on Gaussian process regression for disease progression modeling from time-series of biomarker measurements enabling novel applications beyond the state-of-art, such as the probabilistic prediction of disease staging in testing individuals. Furthermore, the model naturally accounts for missing data, and provides uncertainty quantification of the biomarker evolutions.

Similarly to [5], in this work we focused on the modeling of disease staging represented by a time shift. However, the proposed framework can naturally account for more complex time transformations, provided that a sufficient number of time points is available for each individual.

From the methodological point of view, the proposed model extends current approaches to GP-regression by consistently integrating time-reparameterization and monotonic constraints within a random effect regression framework. Monotonic GPs were introduced in [28] as a principled regularization solution to improve the plausibility of modeling results. For example, the strength of such a regularization approach in biomedical application has been illustrated in survival analysis [15]. Our approach extends this framework by consistently integrating a latent time variable parameter within a random effect model formulation.

The idea of estimating a time transformation in a GP regression framework has been previously used by [18] to account for uncertain measurement times to a microarray dataset of mRNA. However, in that work the estimation of the time uncertainty was subject to a strong prior constraint based on the assumption that the unknown biological time must be similar to the measured one. In the application proposed in our work such an assumption is no longer valid and would ultimately lead to implausible estimations. On the contrary, the proposed GP regression is able to recover the underlying time transformation thanks to the proposed monotonicity regularization.

Finally, thanks to the flexibility of our framework, further extensions of the model will enable to integrate within (**??**) a spatio-temporal covariance model, such as the efficient Kronecker form of [21], to provide a unified framework for jointly modelling time series of images and scalar biomarkers data in a coherent fully Bayesian setting.

4 Conclusions

This work illustrates novel uses of disease progression modeling in the clinical setting. The proposed application shows that disease progression modeling provides a plausible description of the natural history of the disease as well as remarkable diagnostic performances when tested on de-novo individuals. The model used in this study can account for any missing data patterns (longitudinal or across biomarkers), and allows to directly quantify the uncertainty related to the missing information. It thus represents a novel and promising tool for the analysis of clinical trials data.

5 Acknowledgments

EPSRC grants EP/J020990/01 and EP/M020533/1 support DCA and SO's work on this topic. DCA and SO also received support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 666992 (EuroPOND) for this work. MF gratefully acknowledges support from the AXA Research Fund.

References

[1] Randall J Bateman, Chengjie Xiong, Tammie LS Benzinger, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *New England Journal of Medicine*, 367(9):795–804, 2012.

[2] Murat Bilgel, Bruno Jedynak, Dean F Wong, Susan M Resnick, and Jerry L Prince. Temporal trajectory and progression score estimation from voxelwise longitudinal imaging measures: Application to amyloid imaging. In *IPMI*, pages 424–436. Springer, 2015.

[3] Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H Michael Arrighi. Forecasting the global burden of alzheimers disease. *Alzheimer's & dementia*, 3(3):186–191, 2007.

[4] David M Cash, Chris Frost, Leonardo O Iheme, Devrim Ünay, Melek Kandemir, Jurgen Fripp, Olivier Salvado, Pierrick Bourgeat, Martin Reuter, Bruce Fischl, et al. Assessing atrophy measurement techniques in dementia: Results from the miriad atrophy challenge. *NeuroImage*, 123:149–164, 2015.

[5] Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10(5):S400–S410, 2014.

[6] Hubert M Fonteijn, Matthew J Clarkson, Marc Modat, et al. An event-based disease progression model and its application to familial Alzheimer's disease. In *IPMI*, pages 748–759. Springer, 2011.

[7] Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.

[8] Brian A Gordon, Tyler Blazey, Yi Su, Anne M Fagan, David M Holtzman, John C Morris, and Tammie LS Benzinger. Longitudinal β-amyloid deposition and hippocampal volume in preclinical alzheimer disease and suspected non–alzheimer disease pathophysiology. *Jama neurology*, 73(10):1192–1200, 2016.

[9] R Guerrero, A Schmidt-Richberg, C Ledig, T Tong, R Wolz, D Rueckert, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Instantiated mixed effects modeling of alzheimer's disease markers. *NeuroImage*, 142:113–125, 2016.

[10] WJP Henneman, JD Sluimer, J Barnes, WM Van Der Flier, IC Sluimer, NC Fox, Ph Scheltens, H Vrenken, and F Barkhof. Hippocampal atrophy rates in alzheimer disease added value over whole brain volume measures. *Neurology*, 72(11):999–1007, 2009.

[11] Derek LG Hill, Adam J Schwarz, Maria Isaac, Luca Pani, Spiros Vamvakas, Robert Hemmings, Maria C Carrillo, Peng Yu, Jia Sun, Laurel Beckett, et al. Coalition against major diseases/european medicines agency biomarker qualification of hippocampal volume for enrichment of clinical trials in predementia stages of alzheimer's disease. *Alzheimer's & Dementia*, 10(4):421–429, 2014.

[12] Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[13] Clifford R Jack Jr, David S Knopman, Gaël Chételat, Dennis Dickson, Anne M Fagan, Giovanni B Frisoni, William Jagust, Elizabeth C Mormino, Ronald C Petersen, Reisa A Sperling, et al. Suspected non-alzheimer disease pathophysiology [mdash] concept and controversy. *Nature Reviews Neurology*, 2016.

[14] Bruno M Jedynak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T Wyman, David Raunig, C Pierre Jedynak, Brian Caffo, Jerry L Prince, et al. A computational neurodegenerative disease progression score: Method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 63(3):1478–1486, 2012.

[15] Heikki Joensuu, Aki Vehtari, Jaakko Riihimäki, Toshirou Nishida, Sonja E Steigen, Peter Brabec, Lukas Plank, Bengt Nilsson, Claudia Cirilli, Chiara Braconi, et al. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The lancet oncology*, 13(3):265–274, 2012.

[16] Alois Kneip and Theo Gasser. Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics*, pages 82–112, 1988.

[17] Miguel Lázaro-Gredilla, Steven Van Vaerenbergh, and Neil D Lawrence. Overlapping mixtures of gaussian processes for the data association problem. *Pattern Recognition*, 45(4):1386–1395, 2012.

[18] Qiang Liu, Kevin K Lin, Bogi Andersen, Padhraic Smyth, and Alexander Ihler. Estimating replicate time shifts using gaussian process regression. *Bioinformatics*, 26(6):770–776, 2010.

[19] Marco Lorenzi, Michael Donohue, Donata Paternico, Cristina Scarpazza, Susanne Ostrowitzki, Olivier Blin, Elaine Irving, GB Frisoni, Alzheimer's Disease Neuroimaging Initiative, et al. Enrichment through biomarkers in clinical trials of alzheimer's drugs in patients with mild cognitive impairment. *Neurobiology of aging*, 31(8):1443–1451, 2010.

[20] Marco Lorenzi, Ivor J Simpson, Alex F Mendelson, Sjoerd B Vos, M Jorge Cardoso, Marc Modat, Jonathan M Schott, and Sebastien Ourselin. Multimodal image analysis in alzheimers disease via statistical modelling of non-local intensity correlations. *Scientific reports*, 6, 2016.

[21] Marco Lorenzi, Gabriel Ziegler, Daniel C Alexander, and Sebastien Ourselin. Efficient Gaussian process-based modelling and prediction of image time series. In *IPMI*, pages 626–637. Springer, 2015.

[22] Razvan V. Marinescu, Arman Eshaghi, Marco Lorenzi, et al. A vertex clustering model for disease progression: Application to cortical thickness images. In *IPMI*, page to appear. Springer, 2017.

[23] Elizabeth C Mormino, Kathryn V Papp, Dorene M Rentz, Aaron P Schultz, Molly LaPoint, Rebecca Amariglio, Bernard Hanseeuw, Gad A Marshall, Trey Hedden, Keith A Johnson, et al. Heterogeneity in suspected non–alzheimer disease pathophysiology among clinically normal older individuals. *Jama neurology*, 73(10):1185–1191, 2016.

[24] Benson Mwangi, Tian Siva Tian, and Jair C Soares. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244, 2014.

[25] Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078, 2008.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[27] Carl Edward Rasmussen. *Gaussian processes for machine learning*. Springer, 2006.

[28] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *AISTATS*, volume 9, pages 645–652, 2010.

[29] James C Ross and Jennifer G Dy. Nonparametric mixture of gaussian processes with constraints. In *ICML (3)*, pages 1346–1354, 2013.

[30] J-B Schiratti, Stéphanie Allassonniere, Alexandre Routier, Olivier Colliot, and Stanley Durrleman. A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In *IPMI*, pages 564–575. Springer, 2015.

[31] Alexander Schmidt-Richberg, Ricardo Guerrero, Christian Ledig, Helena Molina-Abril, Alejandro F Frangi, Daniel Rueckert, Alzheimers Disease Neuroimaging Initiative, et al. Multi-stage biomarker models for progression estimation in alzheimers disease. In *International Conference on Information Processing in Medical Imaging*, pages 387–398. Springer, 2015.

[32] Robin Wolz, Rolf A Heckemann, Paul Aljabar, Joseph V Hajnal, Alexander Hammers, Jyrki Lötjönen, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Measurement of hippocampal atrophy using 4d graph-cut segmentation: application to adni. *NeuroImage*, 52(1):109–118, 2010.

[33] Zhiyuan Xu, Xiaotong Shen, Wei Pan, Alzheimer's Disease Neuroimaging Initiative, et al. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PloS one*, 9(8):e102312, 2014.

[34] Eric Yang, Michael Farnum, Victor Lobanov, et al. Quantifying the pathophysiological timeline of Alzheimer's disease. *Journal of Alzheimer's Disease*, 26(4):745–753, 2011.

[35] Laurent Younes, Marilyn Albert, Michael I Miller, BIOCARD Research Team, et al. Inferring changepoint times of medial temporal lobe morphometric change in preclinical Alzheimer's disease. *NeuroImage: Clinical*, 5:178–187, 2014.

[36] Alexandra L Young, Neil P Oxtoby, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. A data-driven model of biomarker changes in sporadic alzheimer's disease. *Brain*, 137(9):2564–2577, 2014.

[37] Jonathan Young, Marc Modat, Manuel J Cardoso, Alex Mendelson, Dave Cash, Sebastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.

[38] Peng Yu, Jia Sun, Robin Wolz, Diane Stephenson, James Brewer, Nick C Fox, Patricia E Cole, Clifford R Jack, Derek LG Hill, Adam J Schwarz, et al. Operationalizing hippocampal volume as an enrichment biomarker for amnestic mild cognitive impairment trials: effect of algorithm, test-retest variability, and cut point on trial cost, duration, and sample size. *Neurobiology of aging*, 35(4):808–818, 2014.

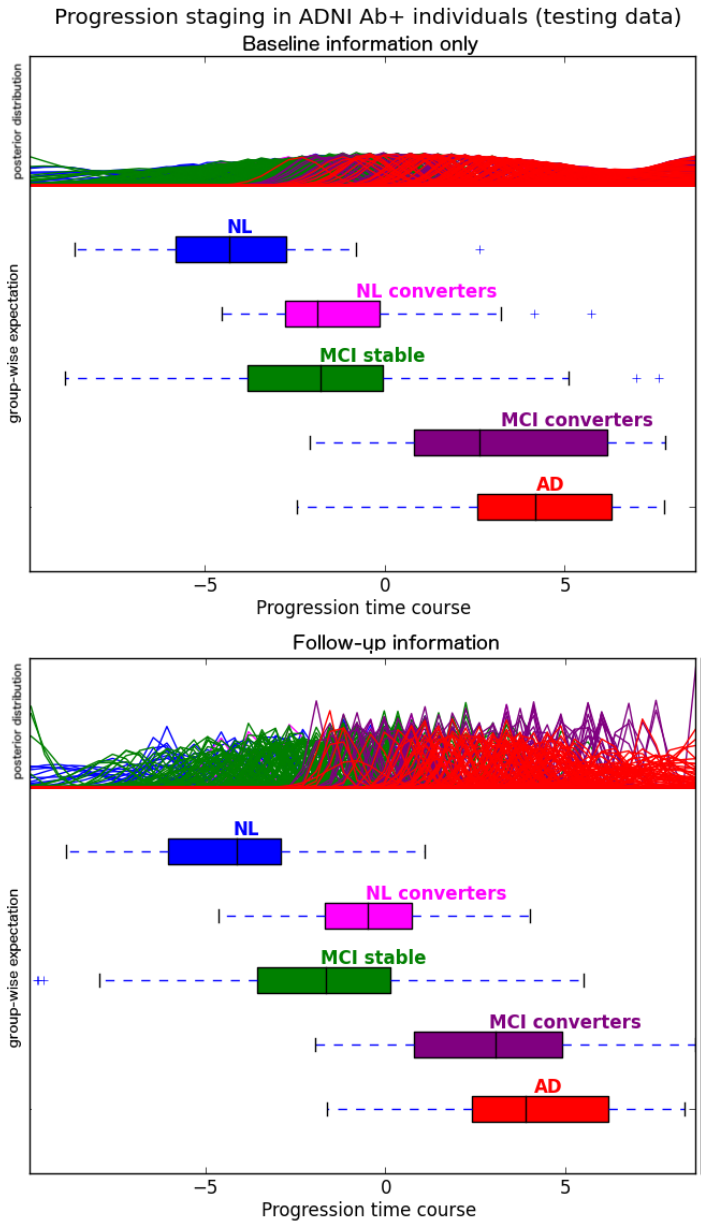


Figure 2: Posterior prediction for the individual time shift in testing data by using i) only the baseline information (top), and ii) the baseline + follow-up information available for each test subject (bottom). The results are similar for both scenarios, although the use of the follow-up information largely reduces the uncertainty of the predictions of the individual pathological stage. Healthy individuals are generally displaced at the early stages of the pathology, while the predictions for MCI and AD patients are associated with respectively intermediate and late progression stages. NL: normal individuals, MCI: mild cognitive impairment, AD: Alzheimer's patients.

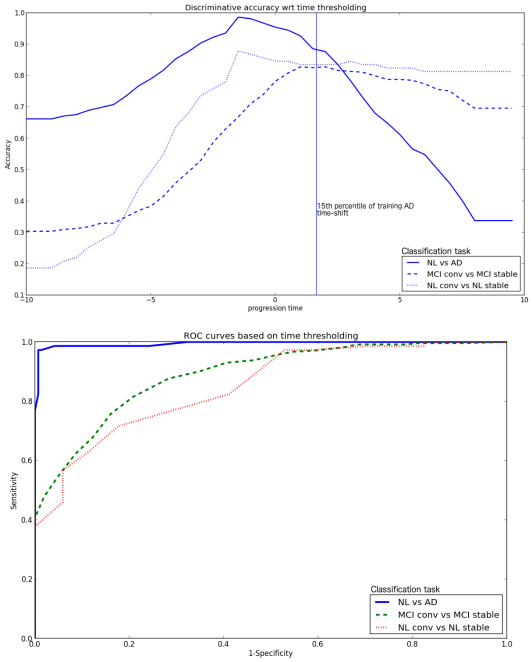
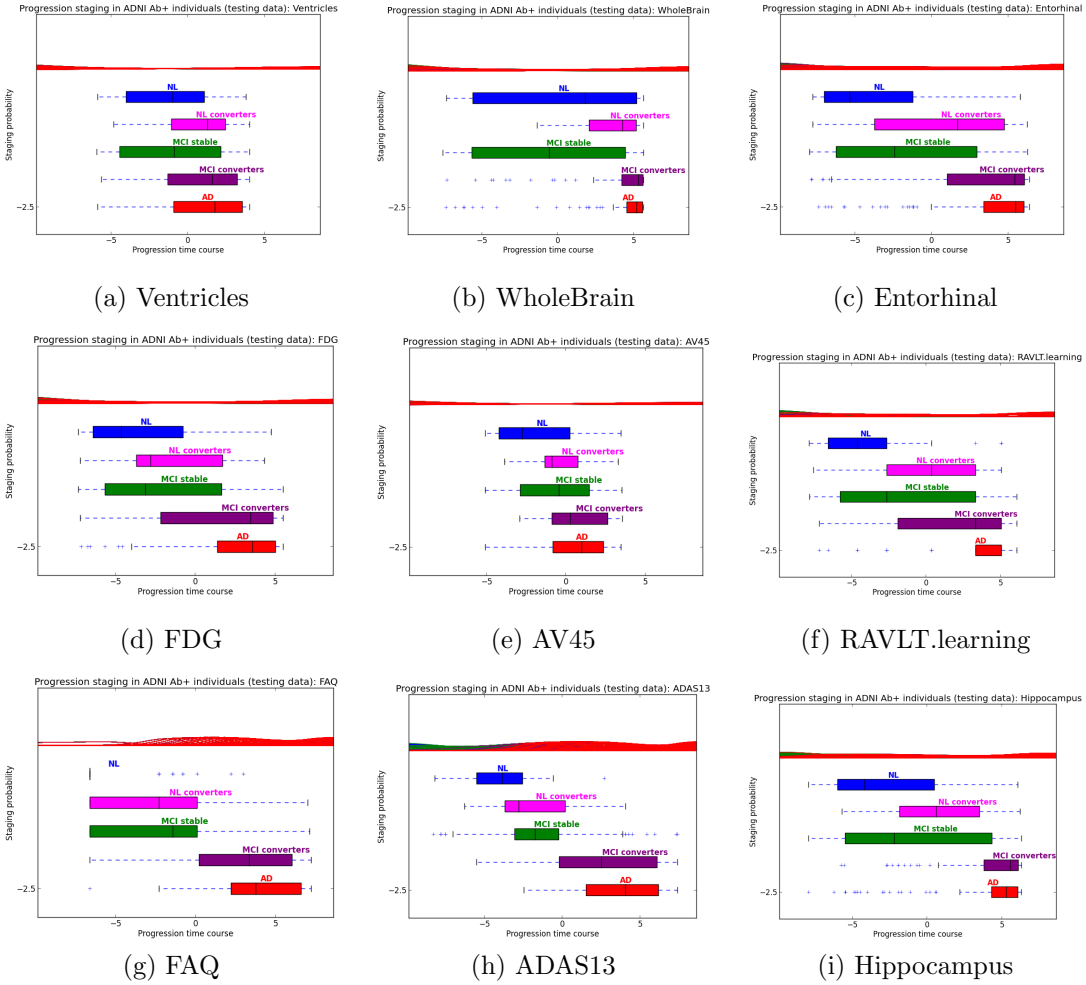


Figure 3: Predictive accuracy of the model when considering the joint set of available biomarkers measurements. The vertical bar indicates the reference threshold value of $t = 1.65$, corresponding to the 15th percentile of the time distribution of the training AD group. MCI: individuals with mild cognitive impairment, AD: Alzheimer's patients.

	Percentiles of the training AD distribution									
	all	Hippo	Ventr	WholeBr	Entor	FDG	AV45	RAVLT	FAQ	ADAS13
NL vs AD (145 vs 74)										
Accuracy	.89	.81	.62	.76	.83	.80	.63	0.82	0.88	0.83
Sensitivity	.83	.84	.52	.9	.82	.74	.82	0.76	0.84	0.75
Specificity	.98	.76	.80	.46	.83	.89	.46	0.94	0.97	0.98
MCI conv vs MCI stable (106 vs 243)										
Accuracy	.82	.67	.62	.69	.7	.71	.69	.67	.79	.79
Sensitivity	.65	.85	.5	.89	.74	.65	.37	.56	.63	.54
Specificity	.90	.59	.68	.60	.68	.73	.75	.71	.86	.9
NL conv vs NL stable (17 vs 74)										
Accuracy	.83	.70	.71	.54	.77	.76	.73	.83	.82	.83
Sensitivity	.18	.47	.41	.82	.52	.29	.27	.35	.17	.17
Specificity	.98	.77	.80	.47	.83	.89	.86	.94	.97	.98

Table 3: Classification results by using the reference time threshold of $t = 1.65$, corresponding to the 15th percentile of the training AD time distribution .



(a) Ventricles (b) WholeBrain (c) Entorhinal

(d) FDG (e) AV45 (f) RAVLT.learning

(g) FAQ (h) ADAS13 (i) Hippocampus

Figure 4: Posterior prediction on testing data by using a single biomarker and the baseline information only.

6 Supplementary Information

6.1 Joint Model: marginal likelihood and inference

Given the sets of individual biomarker measurements $\mathbf{y} = \{(\mathbf{y}^j(t_i))_{i=1}^{k^j}\}_{j=1}^N$, and of D control derivatives $\mathbf{m} = \{m_{b_k}(t'_l)\}_{l=1}^D$ at points $t' = \{t'_l\}_{l=1}^D$ for the progression of each biomarker b_k, the random effect GP model posterior is:

$$\begin{aligned} p\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}, \mathbf{m}\right) &= \frac{1}{Z} p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) p(\mathbf{m}|\dot{\mathbf{f}}) \\ &= p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) \\ &\prod_k \prod_l \Phi\left(\frac{1}{\lambda} \dot{f}_{b_k}(t'_l)\right), \end{aligned} \tag{8}$$

where $\boldsymbol{\nu} = \{\nu^j\}_{j=1}^N$. Thanks to the linearity of GPs under derivation, we have that $Cov\left(\mathbf{f}(t), \dot{\mathbf{f}}(t')\right) = \frac{\mathrm{d}Cov(\mathbf{f}(t), \mathbf{f}(t'))}{\mathrm{d}t'}$, and that the joint distribution $p\left(\mathbf{f}, \dot{\mathbf{f}} | t, t'\right)$ is again a GP

$$p\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | t, t'\right) \quad \sim \quad \mathcal{GP}\left(\mathbf{f}_{joint} | 0, \Sigma_{joint}\right),$$

with $\mathbf{f}_{joint} = \begin{pmatrix} \mathbf{f} \\ \dot{\mathbf{f}} \end{pmatrix}$ distributed as

$$\mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_G(\mathbf{f}(t), \mathbf{f}(t)) & \frac{\partial \Sigma_G(\mathbf{f}(t), \mathbf{f}(t'))}{\partial t'} \\ \frac{\mathrm{d}\Sigma_G(\mathbf{f}(t'), \mathbf{f}(t))}{\mathrm{d}t'} & \frac{\mathrm{d}^2 \Sigma_G(\mathbf{f}(t'), \mathbf{f}(t'))}{\mathrm{d}t'^2} \end{pmatrix}\right].$$

6.1.1 Approximated inference

Due to the non-Gaussianity of the derivative likelihood term, the direct inference on the posterior (8) is not possible due to its analytically intractable form. For this reason, we employ an approximate inference scheme based on classical approaches to Gaussian process with binary activation functions [25]. Following [28], we compute an approximated posterior distribution $q\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}^j, \mathbf{m}\right)$ by replacing the derivative likelihood terms with local un-normalized Gaussian approximations:

$$\begin{aligned} q\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}^j, \mathbf{m}\right) = &\frac{1}{Z_{EP}} p(\mathbf{f}, \dot{\mathbf{f}} | t, t') p(\boldsymbol{\nu}|t) p(\mathbf{y}|\mathbf{f}, \boldsymbol{\nu}) \\ &\prod_k \prod_l \tilde{Z}_{kl} \mathcal{N}(\dot{f}_{b_k}(t'_l) | \tilde{\mu}_{kl}, \tilde{\sigma}^2_{kl}), \end{aligned} \tag{9}$$

where

$$\prod_k \prod_l \tilde{Z}_{kl} \mathcal{N}(\dot{f}_{b_k}(t'_l) | \tilde{\mu}_{kl}, \tilde{\sigma}^2_{kl}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}) \prod_{k,l} \tilde{Z}_{kl},$$

with $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_{kl}]$, and $\tilde{\Sigma}$ is a diagonal matrix with elements $\tilde{\sigma}^2_{kl}$. It follows that the marginal posterior has a Gaussian form, $q\left(\mathbf{f}, \dot{\mathbf{f}}, \boldsymbol{\nu}^j | \mathbf{y}^j, \mathbf{m}\right) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with $\boldsymbol{\mu} = \Sigma\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}_{joint}$, and $\Sigma = (\Sigma^{-1}_{joint} + \tilde{\Sigma}^{-1}_{joint})^{-1}$, where

$$\tilde{\boldsymbol{\mu}}_{joint} = \begin{pmatrix} \mathbf{y} \\ \tilde{\boldsymbol{\mu}} \end{pmatrix}, \quad \text{and} \quad \tilde{\Sigma}_{joint} = \begin{pmatrix} \Sigma_\epsilon + \Sigma_S & 0 \\ 0 & \tilde{\Sigma} \end{pmatrix}.$$

6.1.2 Estimating the EP parameters.

The EP update of the local Gaussian approximation parameters is classically done by iterative moment matching with respect to the product between the cavity distributions $q_{-k'l'}\left(\dot{f}_{b_{k'}}(t'_{l'})\right)$ and the target likelihood term $\Phi\left(\frac{1}{\lambda}\dot{f}_{b_{k'}}(t'_{l'})\right)$.

In the GP case the cavity distribution has a straightforward Gaussian form:

$$\begin{aligned} q_{-k'l'}\left(\dot{f}_{b_{k'}}(t'_{l'})\right) &= \int \prod_{\substack{k\neq k', \\ l\neq l'}} \tilde{Z}_{kl}\mathcal{N}(\dot{f}_{b_k}(t'_l)|\tilde{\mu}_{kl}, \tilde{\sigma}^2_{kl}) d\dot{f}_{b_k}(t'_l) \\ &\sim \mathcal{N}(\dot{f}_{b_{k'}}(t'_{l'})|\mu_{-k'l'}, \sigma_{-k'l'}). \end{aligned} \tag{10}$$

As shown in [28] for univariate monotonic regression, moments and updates of the approximation parameters can be computed in an analogous manner as in the classical GP classification problem [27].

6.1.3 Marginal Likelihood and hyper-parameter estimation

The model's log-marginal likelihood under the EP approximation is:

$$\begin{aligned} \log\mathcal{L} &= -\frac{1}{2}\log|\Sigma_{joint} + \tilde{\Sigma}_{joint}| \\ &\quad -\frac{1}{2}\tilde{\boldsymbol{\mu}}^T_{joint}(\Sigma_{joint} + \tilde{\Sigma}_{joint})^{-1}\tilde{\boldsymbol{\mu}}_{joint} \\ &\quad +\sum_k\sum_l \frac{(\mu_{-kl} - \tilde{\mu}_{kl})^2}{2(\sigma^2_{-kl}) + \tilde{\sigma}^2_{kl})} \\ &\quad +\sum_k\sum_l \log\Phi(\frac{\mu_{-kl}}{\sqrt{\lambda^2_k + \sigma^2_{-kl}}}) \\ &\quad +\frac{1}{2}\sum_k\sum_l \log(\sigma^2_{-kl} + \tilde{\sigma}^2_{kl}). \end{aligned} \tag{11}$$

In what follows, the optimal parameters are obtained by maximising $\log\mathcal{L}$ through conjugate gradient descent, via alternate optimization between the hyper-parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}^j_G$, and the individuals' time-shifts d^j. The position of the derivative points was updated at each iteration, according to the changes of the GP domain. Regularisation was also enforced by introducing

Gaussian priors for the parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_G^j$. We note that the block structure of the GP covariance allows the computation of the gradients with respect to the biomarkers' and individual parameters by working on matrices of much smaller dimension than the one of the whole GP, thus considerably improving the numerical stability and the computational efficiency of the optimization procedure.

		$N = 20$				$N = 100$			
		σ				σ			
		0.1	0.2	0.3	0.4	0.1	0.2	0.3	.4
N_b	4	.95 (.03)	.86 (.08)	.71 (.17)	.46 (.29)	.91 (.04)	.89(.04)	.76 (.17)	.75 (.12)
	8	.97 (.01)	.91 (.06)	.86 (.06)	.66 (.3)	.94 (.04)	.94 (.02)	.88 (.06)	.84 (.07)

Table 4: Mean (sd) R^2 correlation coefficient across folds between estimated individual time-shifts and ground truth time reference.

6.2 Model benchmarking on synthetic multivariate progressions

We benchmarked the model with respect to synthetic multivariate biomarker progressions. We generated random multivariate sigmoid functions for N_b biomarkers, $\mathbf{f}(\tau) = (f_{b_1}(\tau), f_{b_2}(\tau), \ldots, f_{b_{N_b}}(\tau))^\top$, with $f_{b_k}(\tau) = 1/(1 + \exp(-\alpha_k \tau))$, $\tau \in [0, 15]$ and $\alpha_k \sim \mathcal{N}(0, .06)$, and we sampled N individual noisy trajectories at time points τ_k^j: $\mathbf{y}_k^j(\tau_k^j) = f_k(\tau_k^j) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For each individual we used the same initial sampling time point for every biomarker, while the number of samples per biomarker was allowed to independently vary between 1 and 4. The individual time points were subsequently centered by their mean μ_k^j to obtain shifted time-points $t_k^j = \tau_k^j - \mu_k^j$ defined in the interval $[-2, 2]$.

The model was applied to estimate biomarker progressions and individual time-shifts with respect to different combinations of trajectory noise σ, sample size N, and number of biomarkers N_b. The accuracy of the model in reconstructing the original time series was quantified by Pearson's correlation between the estimated time-shift d^j and the original individual time reference. The experiments were repeated 10 times for each configuration of parameters $\sigma \in \{0.1, 0.2, 0.3, 0.4\}$, $N_b \in \{4, 8\}$, and $N \in \{20, 100\}$.

6.2.1 Results.

Table 4 reports summary correlations between time-shift estimation and the ground truth individual sampling time. The correlation values are generally high, and increase with lower noise levels. Interestingly, the increase in number of modelled biomarkers is associated with a better performance in recovering the underlying disease staging. We also observe that larger sample sizes are associated with higher correlation values, especially with increasing noise levels. We note however an exception for the case $\sigma = 0.1$ where, although the overall performance is still high, the correlation slightly decreases with $N = 100$.

6.2.2 Model benchmarking with respect to grace

The R package GRACE (v 1.0) was used to estimate the multivariate biomarker progression curves from the training set used in this study.

Figure 5 shows the relationship between the estimated individual time-shift. Although the time range estimated by the GP model is roughly double with the respect to the GRACE one, there is a strong agreement between the *relative* positioning of the training individuals along the disease trajectory.

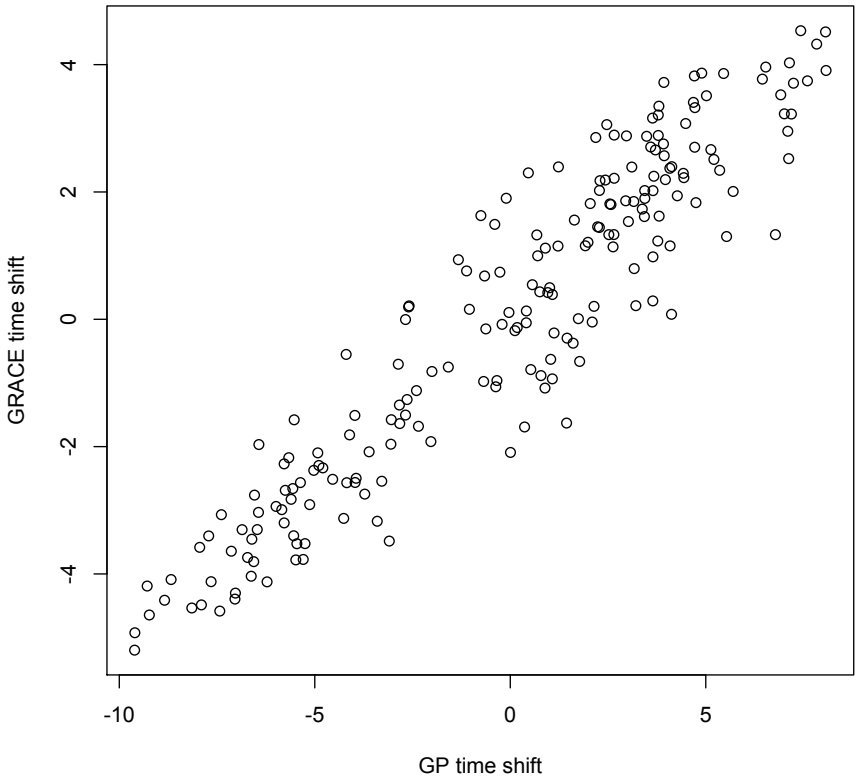
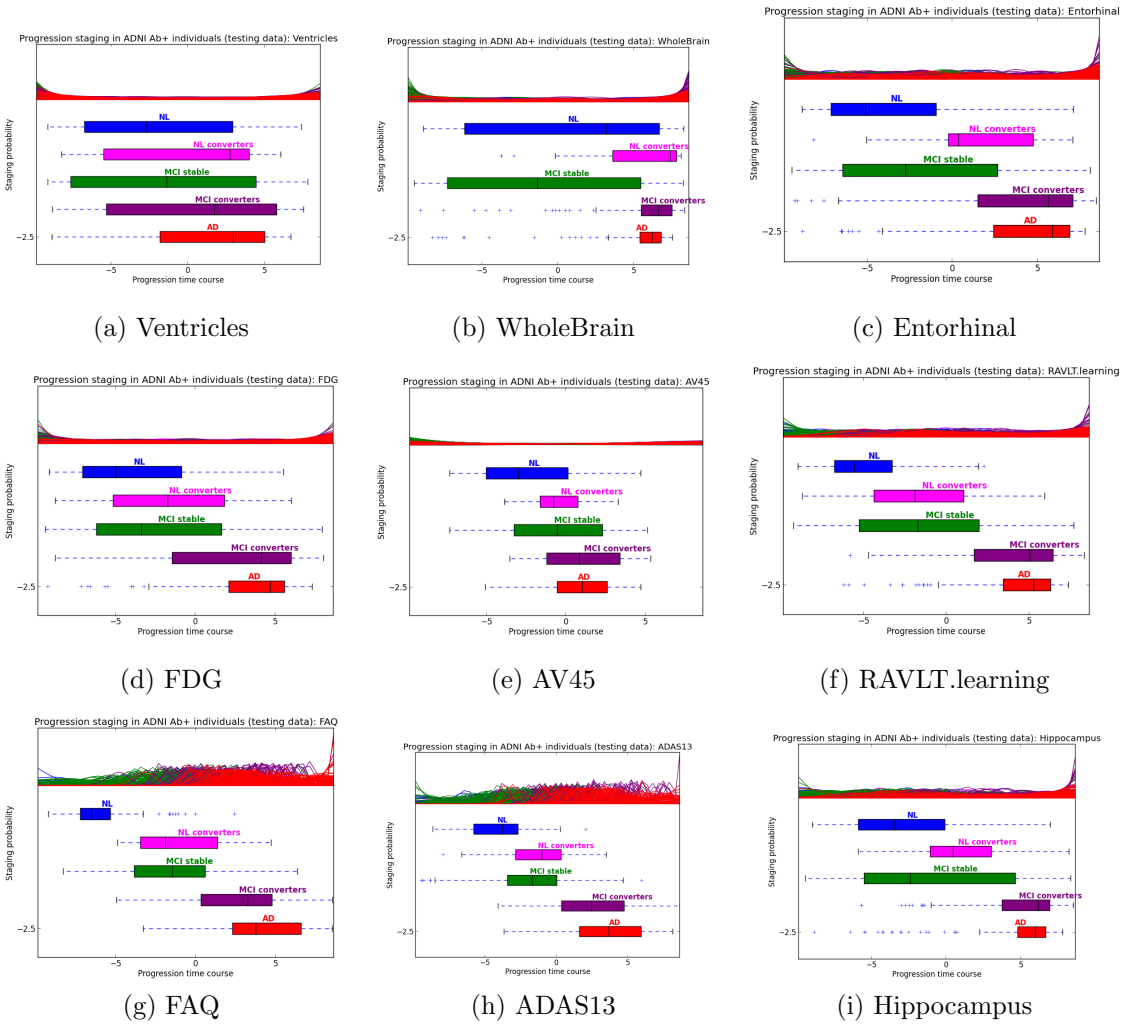


Figure 5: Comparison between the shift estimated with our GP progression model (x-axis), and the one estimated by GRACE (y-axis). Although the time range estimated by the GP model is roughly double with the respect to the GRACE one, there is a strong agreement between the *relative* positioning of the training individuals along the disease trajectory.

6.2.3 Supplementary Figure



(a) Ventricles (b) WholeBrain (c) Entorhinal

(d) FDG (e) AV45 (f) RAVLT.learning

(g) FAQ (h) ADAS13 (i) Hippocampus

Figure 6: Posterior prediction on testing data by using a single biomarker and the follow-up information.

6.2.4 Data preparation

```
library("ADNIMERGE")
ridAD = unique(subset(adnimerge,DX=="Dementia")$RID)
ridNL = unique(subset(adnimerge,DX=="NL")$RID)
ridMCI = unique(subset(adnimerge,DX=="MCI")$RID)

ADreverted = c(167, 1226, 4641)
ridAD = ridAD[!ridAD%in%ADreverted]

NLconverted = c(15, 22, 35, 55, 61, 106, 112, 127, 156, 171, 210, 223, 232,
259, 420, 454, 459, 467, 520, 545, 548, 555, 558, 602, 605, 622, 680, 722,
778, 779, 842, 843, 883, 899, 920, 972, 985, 1063, 1123, 1169, 1190, 1194,
1200, 1202, 1203, 2150, 4041, 4071, 4092, 4218, 4262, 4385, 4474, 4506, 4566,
4577, 4579, 4652, 4855, 5096, 5121, 5207, 5273)
ridNL = ridNL[!ridNL%in%NLconverted]


ridConv = subset(adnimerge,RID%in%ridMCI&DX=="MCI to Dementia")$RID
ridReverter = subset(adnimerge,RID%in%ridConv&DX=="Dementia to MCI")$RID
ridReverter = c(429, 4706)
ridConv = ridConv[!ridConv%in%ridReverter]

ridMCI = c(ridMCI,ADreverted)
ridNConv = ridMCI[which(!ridMCI%in%ridConv)]
ridNConv =  ridNConv[!ridNConv%in%c(ridConv,ridAD,ridNL,NLconverted)]

ridAD = ridAD[-which(ridAD%in%ridConv)]

Abpos = read.csv("AbposADNI.csv",skip=1)
ridABpos = Abpos$RID

Set = subset(adnimerge,RID%in%c(ridNConv,ridConv,ridAD,ridNL, NLconverted),
select=c("RID","Month","DX","Hippocampus",
"Ventricles","WholeBrain","Entorhinal","FDG","AV45",
"RAVLT.learning","FAQ", "ADAS13","ICV.bl"))
Set$Hippocampus = Set$Hippocampus/Set$ICV.bl
Set$WholeBrain = Set$WholeBrain/Set$ICV.bl
Set$Entorhinal = Set$Entorhinal/Set$ICV.bl
Set$Ventricles = Set$Ventricles/Set$ICV.bl

Set = subset(Set,select=c("RID","Month","DX","Hippocampus","Ventricles",
"WholeBrain","Entorhinal","FDG","AV45", "RAVLT.learning","FAQ", "ADAS13"))

RIDnoNA = subset(Set,Month==0)$RID[which(apply(is.na(subset(Set,Month==0)),
1,any)==FALSE)]
```

```
SetnoNA = subset(Set,RID%in%ridABpos&RID%in%RIDnoNA&RID%in%
c(ridConv,ridAD,ridNL,NLconverted))
trainRID = sample(unique(SetnoNA$RID),200)
trainSet = subset(Set,RID%in%trainRID)
testSet = subset(Set,!RID%in%trainRID&RID%in%Abpos$RID)

newSet = trainSet

for (i in seq(4,length(names(newSet))))
    {
newSet[,i] = rank(newSet[,i],na.last='keep')/
length(newSet[,i][which(!is.na(newSet[,i]))])
}



newSet_test = testSet

for (i in seq(4,length(names(newSet)))){
for (j in seq(1,length(newSet_test[,i])))
{
if (!is.na(testSet[j,i]))
{
newSet_test[j,i] = rank(c(testSet[j,i],trainSet[,i]),
na.last='keep')[1]/
(length(trainSet[,i][which(!is.na(trainSet[,i]))])+1)
}

}
}



newSet$FDG = 1-newSet$FDG
newSet$Hippocampus = 1-newSet$Hippocampus
newSet$WholeBrain = 1-newSet$WholeBrain
newSet$Entorhinal = 1-newSet$Entorhinal
newSet$RAVLT.learning = 1 - newSet$RAVLT.learning

newSet_test$FDG = 1-newSet_test$FDG
newSet_test$Hippocampus = 1-newSet_test$Hippocampus
newSet_test$WholeBrain = 1-newSet_test$WholeBrain
newSet_test$Entorhinal = 1-newSet_test$Entorhinal
newSet_test$RAVLT.learning = 1 - newSet_test$RAVLT.learning
```

```
write.csv(newSet,"ADNIDataTrain.csv")
write.csv(newSet_test,"ADNIDataTest.csv")

write.csv(ridAD,"ridAD.csv",row.names=FALSE)
write.csv(NLconverted,"ridNLconverted.csv",row.names=FALSE)
write.csv(ridConv,"ridConv.csv",row.names=FALSE)
write.csv(ridNL,"ridNL.csv",row.names=FALSE)
write.csv(ridNConv[ridNConv%in%ridABpos],"ridNConv.csv",row.names=FALSE)
```

Efficient Gaussian Process-Based Modelling and Prediction of Image Time Series

Marco Lorenzi[1], Gabriel Ziegler[2], Daniel C. Alexander, and Sebastien Ourselin[1]

1. Translational Imaging Group, CMIC, UCL, London, UK
2. Wellcome Trust Centre for Neuroimaging, UCL, London, UK
3. Centre for Medical Image Computing, CMIC, UCL, London, UK

Originally published in:
Information Processing in Medical Imaging (IPMI), 24:626-37, Springer, LNCS, 2015

Abstract

In this work we propose a novel Gaussian process-based spatio-temporal model of time series of images. By assuming separability of spatial and temporal processes we provide a very efficient and robust formulation for the marginal likelihood computation and the posterior prediction. The model adaptively accounts for local spatial correlations of the data, and the covariance structure is effectively parameterised by the Kronecker product of covariance matrices of very small size, each encoding only a single direction in space. We provide a simple and flexible framework for within- and between-subject modelling and prediction. In particular, we introduce the Hoffman-Ribak method for efficient inference on posterior processes and its uncertainty. The proposed framework is applied in the context of longitudinal modelling in Alzheimer's disease.We firstly demonstrate the advantage of our non-parametric method for modelling of within-subject structural changes.The results show that non-parametric methods demonstrably outperform conventional parametric methods. Then the framework is extended to optimize complex parametrized covariate kernels. Using Bayesian model comparison via marginal likelihood the framework enables to compare different hypotheses about individual change processes of images.

1 Introduction

Modelling longitudinal changes in organs is fundamental for the understanding of biological and pathological processes. For instance the development of a spatio-temporal model of disease progression in Alzheimer's disease (AD) from time series of magnetic resonance images (MRIs) would be highly valuable for the fundamental understanding of the disease process, for diagnostic purposes and individual predictions, and for testing the efficacy of disease modifying drugs in clinical trials.

The consistent modelling and prediction of spatio-temporal changes in longitudinal MRI is still an important challenge from both methodological and computational perspectives. In fact, flexible modelling instruments are required in order to robustly capture meaningful pathological accelerations specific to sensitive brain regions. Moreover, since a biological model of local brain changes is often unknown, it is important to develop optimal models in terms of statistical complexity.

Many of the previous works on spatio-temporal modelling of image time series are based on non-linear image registration, describing signal differences between images as local spatial transformations [1, 2, 3, 4]. However, statistical inference in registration models is often limited, due to the computational complexity, and since image-registration is generally not flexible enough to perform model comparisons and clinical prediction, to account for covariates and for the within- and between subjects heterogeneity.

A statistical focus on the modeling of image time series is commonly provided by parametric linear modelling frameworks (GLM) [5]. However, GLM approaches are often limited by the choice of arbitrary model complexity and spatial resolution at which the data is analyzed. Even though flexible non-parametric models have been proposed for the analysis of spatio-temporal signals in brain images [6, 7], their computational complexity still prevents the straightforward application in time series of high-resolution MRIs. Non-parametric Gaussian process (GP) models have emerged as a flexible and elegant Bayesian approach for prediction and modelling in manifold applications [8], and have been recently successfully introduced to the field of neuroimaging, e.g. in the context of single-case inference in aging [9]. However, the application of GPs to the voxel-wise modelling of image time series is to date very challenging, since the specification of the joint covariance structure of the image features is in general computationally prohibitive.

In this work we introduce a generative model of spatio-temporal changes based on GPs, to provide a flexible and computationally efficient approach to the analysis of aligned image time series by accounting for spatial and temporal correlation. In particular, by assuming a local spatial correlation model and the separability between spatial and temporal changes, we introduce a very efficient formulation based on a covariance structure parameterized by the Kronecker product of small size covariance matrices [10]. The proposed model extends GLM approaches by providing a flexible and efficient statistical tool for the analysis of image features from spatially aligned time series, for instance by allowing statistical inference on the model parameters.

The paper is organized as follows. In Section 2 we propose our generative model of longitudinal changes in image time series, while in Section 3 we provide computationally tractable optimization and prediction schemes. We also introduce a novel computational scheme based on the Hoffman-Ribak method for the statistical inference in high dimensional GP-based spatio-temporal models. Finally, in Sections 4 and 5, we apply the model in the context of longitudinal data from from the Alzheimer's Disease Neuroimaging Initiative (ADNI) for 1) within-subject modelling and prediction of local and regional brain longitudinal changes, and 2) group-wise joint modelling of local ventricle growth rates based on socio-demographics, genetic factors, and clinical scores.

2 A generative model for within-subject image time series

Let $u = (x, y, z)$ be the 3-dimensional spatial coordinate system and t the temporal dimension. We consider the image time series I as a discretely sampled spatio-temporal signal of dimensions $N \times N \times N \times N_T$, where N is the dimension of the sampling grid on a single spatial axis, and N_T is the number of time points[1]. In the following sections we represent the image time-series I as a single dimensional array of dimensions $N^3 N_T$. We model the image time series $I(u, t)$ as a

[1]For simplicity we focus on an even sampling across spatial directions, even though the generalization of the proposed model to the uneven case is straightforward.

realization of a latent spatio-temporal process $f(u,t)$ with additive noise:

$$I(u,t) = f(u,t) + \epsilon(u,t) \ . \tag{1}$$

The true signal will be modelled as a GP with zero mean and covariance Σ, while ϵ is assumed to be i.i.d. Gaussian distributed measurement noise $\epsilon(u,t) \sim \mathcal{N}(0,\sigma^2)$. Here we first assume that spatial and temporal processes are *separable*, and thus that the covariance matrix Σ can be factorised in the Kronecker product of independent spatial and temporal covariance matrices: $\Sigma = \Sigma_S \otimes \Sigma_T$.

This is a valid modeling assumption when the temporal properties of the signal are similar across space; for instance, when analyzing within-subject time series of brain MRIs in AD the expected pathological change rates are generally mild and slowly varying across the brain. Second, a central assumption made in this paper is that the spatial dependencies of the signal are *local*, i.e. that the image intensities are smoothly varying and correlated within a spatial neighborhood of radius l_s. We note that our assumptions about separability and stationarity are compatible with the spatio-temporal correlation models commonly assumed by registration-based approaches.

A reasonable choice for such a local spatial covariance structure is a negative squared exponential model $\Sigma_S(u_1,u_2) = \lambda_s \exp(-\frac{\|u_1-u_2\|^2}{2l_s})$, where λ_s is the global spatial amplitude parameter, and l_s is the length-scale of the Gaussian spatial neighborhood. We observe that such a covariance structure is *stationary* with respect to the space parameters. Furthermore we can exploit the separability properties of the negative exponential function to note that given two separate spatial locations $u_1 = (x_1, y_1, z_1)$ and $u_2 = (x_2, y_2, z_2)$ we have

$$\Sigma_S(u_1,u_2) = \lambda_s \exp(-\frac{(x_1-x_2)^2}{2l_s}) \exp(-\frac{(y_1-y_2)^2}{2l_s}) \exp(-\frac{(z_1-z_2)^2}{2l_s}) \ .$$

For this reason the covariance matrix Σ_S can be further decomposed as the Kronecker product of covariance matrices of 1-dimensional processes: $\Sigma = K_x \otimes K_y \otimes K_z \otimes \Sigma_T$. We observe that the model is here conveniently represented by the product of independent covariances of significantly smaller size, and is completely identified by the spatial, temporal and noise parameters. In particular the proposed model is flexible with respect to the temporal covariance matrix Σ_T, which can be expressed in terms of complex mixed-effects structure, and can account for covariates and different progression models. For instance, in this work the matrix Σ_T is first specified in order to model the temporal progression observed in time series of images (Section 4), and then is used to model the influence of anatomical, genetic, clinical, and sociodemographic covariates on individual atrophy rates modelled by non-linear registration (Section 5).

3 Inference in Gaussian processes with Kronecker structure

The GP-based generative model with Kronecker covariance structure outlined in this work provides a powerful and efficient framework for prediction using image time series. Here we provide the main results concerning the marginal likelihood computation, the hyper-parameter optimization and the posterior prediction.

Let $(U_{K_x}, S_{K_x} = \text{diag}(\lambda_1^x, \ldots, \lambda_N^x))$ and $(U_T, S_T = \text{diag}(\lambda_1^t, \ldots, \lambda_{N_T}^t))$ be the eigenvectors and eigenvalues associated to the one-dimensional spatial and temporal covariance matrices K_x and Σ_T. This eigendecomposition problem can be easily and efficiently solved beforehand offline. We further introduce the shortform notation $\bigotimes A = A_x \otimes A_y \otimes A_z$.

3.0.1 Log-Marginal Likelihood.

The marginal likelihood of the model (1) is the following:

$$\log \mathcal{L} = -\frac{1}{2}\sum_{i,j,k,l} \log(\lambda_i^x \lambda_j^y \lambda_k^z \lambda_l^t + \sigma^2) - \frac{1}{2} {V_I}^T (\bigotimes S_K \otimes S_T + \sigma^2 Id)^{-1} V_I + const \ , \tag{2}$$

with $const = -\frac{N^3 N_T}{2}\log(2\pi)$, $V_I = \text{vec}\left[(U_{K_z}^T \otimes U_T^T)\tilde{I}(U_{K_x} \otimes U_{K_y})\right]$, and where $\tilde{I}$ is the matricization of I into a 2 dimensional matrix of dimension $N^2 \times NN_T$, and $\lambda_i^x, \lambda_j^y, \lambda_k^z$ and λ_l^t are the eigenvalues of respectively K_x, K_y, K_z and Σ_t. The computation of the vector V_I requires the storage and multiplication of matrices of relatively small sizes, respectively $N^2 \times N^2$, $N^2 \times NN_T$ and $NN_T \times NN_T$. The product $(\bigotimes S_K \otimes S_T + \sigma^2 Id)^{-1} V_I$ can be finally computed as the solution of the linear system $(\bigotimes S_K \otimes S_T + \sigma^2 Id)\mathbf{X} = V_I$, which is straightforward since $(\bigotimes S_K \otimes S_T + \sigma^2 Id)$ is diagonal.

3.0.2 Hyperparameter optimization.

The derivative of the log-likelihood (2) with respect to the model parameters $\boldsymbol{\theta}$ is:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \log \mathcal{L} = -\frac{1}{2} Tr\left((\bigotimes K \otimes \Sigma_T + \sigma^2 Id)^{-1} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} (\bigotimes K \otimes \Sigma_T + \sigma^2 Id) \right) \\ - \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} I^T (\bigotimes K \otimes \Sigma_T + \sigma^2 Id)^{-1} I \ . \end{aligned} \tag{3}$$

It can be shown that formula (3) can be efficiently computed with respect to each model parameters. For instance, the gradient with respect to the noise parameter can be expressed in the form:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} \log \mathcal{L} = -\frac{1}{2} \sum_{i,j,k,l} (\lambda_i^x \lambda_j^y \lambda_k^z \lambda_l^t + \sigma^2)^{-1} + \frac{1}{2} {V_I}^T (\bigotimes S_K \otimes S_T + \sigma^2 Id)^{-2} V_I \ . \tag{4}$$

3.0.3 Prediction.

A major strength of a GP framework for image time series is that it easily enables probabilistic predictions based on given observations. The proposed generative model allows us to consider the predictive distributions of the latent spatio-temporal process at any testing locations u^* and timepoints t^*. Given image time series $I(u,t)$, we now aim at predicting the image I^* at $N^* \times N_T^*$ testing coordinates $\{u^*, t^*\}$. Let us define $\Sigma_{I,I^*} = \Sigma(u,t,u^*,t^*)$ the cross-covariance matrix of training and testing data, and $\Sigma_{I^*,I^*} = \Sigma(u^*,t^*,u^*,t^*)$ the covariance evaluated on the new coordinates. The joint GP model of training and testing data is:

$$\begin{pmatrix} I(u,t) \\ I^*(u^*,t^*) \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma + \sigma^2 Id & \Sigma_{I,I^*} \\ \Sigma_{I^*,I} & \Sigma_{I^*,I^*} + \sigma^2 Id \end{pmatrix} \right] \ , \tag{5}$$

and it can be easily shown that the posterior distribution of I^* conditioned on the observed time series I and parameters $\boldsymbol{\theta}$ is [8]:

$$\begin{aligned} I^* | I, \{u^*, t^*\}, \boldsymbol{\theta} \sim & \mathcal{N}\Big(\mu^*, \Sigma^*\Big), \text{ where } \mu^* = \Sigma_{I,I^*}\Sigma^{-1} I \\ & \text{and } \Sigma^* = \Sigma_{I^*,I^*} - \Sigma_{I,I^*}\Sigma^{-1}\Sigma_{I^*,I} + \sigma^2 Id \ . \end{aligned} \tag{6}$$

From the practical perspective, we notice that by definition the new covariance matrices still have a Kronecker product form: $\Sigma_{I,I^*} = K_{x,x^*} \otimes K_{y,y^*} \otimes K_{z,z^*} \otimes \Sigma_{t,t^*}$, and $\Sigma_{I^*,I^*} = K_{x^*,x^*} \otimes K_{y^*,y^*} \otimes K_{z^*,z^*} \otimes \Sigma_{t^*,t^*}$. The predicted mean μ^* at coordinates $\{u^*, t^*\}$ is then

$$\mu^* = \left(K_{x,x^*} U_{K_x} \otimes K_{y,y^*} U_{K_y} \otimes K_{z,z^*} U_{K_z} \otimes \Sigma_{t,t^*} U_T\right) \left(\bigotimes S_K \otimes S_T + \sigma^2 Id\right)^{-1} V_I \ ,$$

which can be computed efficiently by noting that the matrix to be inverted is diagonal and by using the product rule of the Kronecker operator. While the posterior form (6) can also be used to evaluate the posterior marginal covariance, certain considerations are necessary for a tractable approach. Indeed, the covariance matrix Σ^* is computed from Σ, Σ_{I^*,I^*} and $\Sigma_{I^*,I}$, which are evaluated on different sets of spatial and temporal coordinates. In particular, the Kronecker structure is lost and in the absence of further assumptions the matrix Σ^* must therefore be explicitly computed, generally leading to impractical solutions.

3.0.4 Hoffman-Ribak method for posterior sampling.

We propose to compute the *sample distribution* of (6) using the Hoffman-Ribak method (HR) introduced in the late 1990s in the astrophysics literature [11]. Given the Gaussian distribution (5) partitioned into training (observed) and testing (unobserved) components, the HR method provides a computationally efficient and exact algorithm for sampling from (6) consisting of the following two steps:

- Sample a random observation (Y, Y^*) from the joint distribution (5),
- Compute a sample Z of the marginal posterior (6) according to $Z = Y^* + \Sigma_{I^*,I}(\Sigma + \sigma^2 Id)^{-1} Y$.

Despite its simple formulation, the HR method cannot be straightforwardly applied in our case as sampling from the very high dimensional joint distribution is generally prohibitive. Therefore, instead of focusing on predicting time series at arbitrary spatial and temporal coordinates, we provide here an efficient scheme for spatio-temporal prediction at arbitrary time points $T^* = \{t^*\}$ evaluated in the same spatial coordinates of the training image time-series I. Under this assumption the matrices Σ, Σ_{I^*,I^*} and $\Sigma_{I^*,I}$ differ in the temporal part only,

$$\Sigma = \Sigma_S \otimes \Sigma_T + \sigma^2 Id; \quad \Sigma_{I^*,I^*} = \Sigma_S \otimes \Sigma_{T^*,T^*}; \quad \Sigma_{I^*,I} = \Sigma_S \otimes \Sigma_{T^*,T} + \sigma^2 Id,$$

and it is simple to show that the joint covariance is $\Sigma^{joint} = P(\Sigma_S \otimes \Sigma_{T^j} + \sigma^2 Id)P^T$, where P is a structured permutation matrix, and $\Sigma_{T^j} = \left(\begin{smallmatrix} \Sigma_T & \Sigma_{t,t^*} \\ \Sigma_{t^*,t} & \Sigma_{t^*,t^*} \end{smallmatrix}\right)$. A sample Z from the joint distribution can thus be easily computed as $Z = P(U\Lambda)X$, where X is a standard multivariate normal distributed vector, and $U\Lambda^2 U^T$ is the eigen-decomposition of the covariance $(\Sigma_S \otimes \Sigma_{T^j} + \sigma^2 Id)$. Eigen-decomposition and matrix multiplication can be efficiently computed by virtue of the properties of the Kronecker product.

In the following sections, after validating the proposed framework in a controlled setting, we provide a modelling application in the context of longitudinal modelling in AD.

4 Model Validation

4.0.1 Estimation of the Spatio-temporal Properties in Synthetic Data.

Here, we test the ability of the proposed GP model to correctly estimate the underlying spatial and temporal properties prescribed in synthetic data. We chose a time-series of brain MRIs composed of 6 aligned longitudinal gray matter (GM) segment images of an example ADNI patient, and we applied Gaussian smoothing to obtain synthetic samples of a spatio-temporal process with predefined spatial correlation and signal to noise ratio. Moreover we generated synthetic longitudinal progressions of increasing temporal complexity following respectively voxel-wise linear, quadratic and cubic functions of time estimated through a general linear model (GLM). Furthermore, longitudinal changes in the synthetic time series were modelled with the proposed GP model. We applied a squared exponential model for the temporal covariance parameterized by the temporal length-scale l_t. A maximum-a-posteriori (MAP) estimate of the parameters was obtained by using Gauss-Newton optimization scheme of the log-hyperparameters, using multivariate uninformative Gaussian hyperprior with log-hyperparameters $\mu_h = [-2, -2, 0, 3]$ and $\Sigma_h = \text{diag}([5, 5, 1, 5])$ for respectively $(\sigma^2, l_s, \lambda_s, l_t)$.

Table 1 shows the relationship between the spatio-temporal properties of the synthetic data and the MAP estimates of the GP parameters. Noticeably, the estimated spatial length-scale closely resembles the global smoothness parameter of the synthetic data, adaptively accounting for image smoothness properties. Additionally, we observed that the estimated temporal length-scale decreased when modeling longitudinal progressions of higher order models. Thus, the model also correctly denotes the increased complexity of the temporal changes.

Spatial smoothness (mm)	l_s	σ^2	λ_s
0	0.09	9e-6	0.7
0.5	0.81	5e-6	0.64
1	1.2	3e-6	0.53
2	2.3	1e-6	0.5
3	3.3	3e-10	0.48
4	4.3	7e-11	0.47

Temporal progression	l_t (log-values)
linear	4.3
quadratic	1.79
cubic	1.72

Table 1: Estimation of the global spatial and temporal properties. The estimated spatial length-scale l_s closely correspond to the global smoothness of the synthetic data, while the noise term and the signal amplitude decrease with increasingly smoother data. The estimated temporal length-scale is inversely proportional to the underlying complexity of the temporal progression.

4.0.2 Within-Subject Modelling and Prediction of Longitudinal Changes.

We chose high-resolution longitudinal images of 10 AD patients, 10 patients with mild cognitive impairment (MCIc) subsequently converting to AD, and 10 healthy controls from the ADNI dataset. AD patients and healthy controls (HC) had 4 images per participant, corresponding to baseline, 6 months, 1 and 2 years scans, while for MCIc patients additional images corresponding to 3 or 4 years were available. The images were processed according to established procedures consisting of joint bias correction, tissue segmentation, alignment to the within-subject average anatomy, and

non-linear normalization to a group-wise anatomical reference [12]. The final image size was of 100^3 cubic voxels with isotropic resolution of $1.5\,mm$.

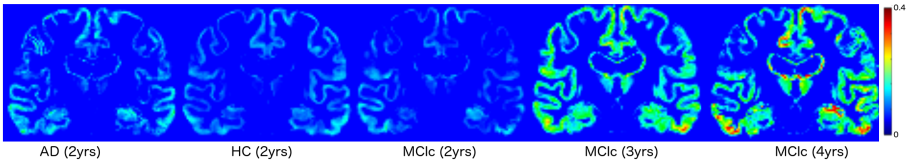


Figure 1: Group-wise average absolute differences between extrapolated images and real ones. The GP model was trained on scans from 3 time points corresponding to baseline, 6 months and 1 year. Errors were generally found to be proportional to the extrapolation time.

The longitudinal changes in the resulting time series of processed gray matter density maps were modelled according to the proposed GP model. The model was estimated for each subject by using 3 training images corresponding to baseline, 6 months and 1 year scans. In order to capture meaningful non-linear trends during disease progression to AD, we also applied the GP model in the MCIc group by using 4 and 5 training images, corresponding to the time range from baseline to respectively 2 and 3 years follow-up.

We applied the optimization scheme illustrated in Section 4.0.1 while imposing an informative prior on the temporal length-scale parameter with log-mean and -variance of 3 and 0.1 respectively. This choice was done in virtue of the experimental results illustrated in Table 1 in order to promote a moderately non-linear behaviour of the GP model, and at the same time avoid overfitting on the limited number of within-subject observations. The resulting computational time for the parameter estimation was of about 5 minutes per subject on a standard PC (with 2.6 GHz, QuadCore, 16GB RAM). The predictive accuracy of the model was then tested by voxel-wise comparison of the extrapolated image series with respect to the corresponding ground truth follow-up images, and compared with respect to a standard linear and quadratic voxel-by-voxel model using within-subject GLM. The group-wise average voxel-wise absolute differences between extrapolated images and real ones are shown in Figure 1. Errors were generally found to be proportional to the extrapolation time. Table 2 shows that the results of the GP model are comparable to those obtained by linear modelling when training on 3 time points only. However, the prediction of the GP model significantly improves the linear one when using more training points. This result

	AD	HC	MCI	
N train points	3	3	4	5
GP	1.9	1.9	2.9*	2.5*
GLM linear	1.9	2	3.1	2.7
GLM quadratic	6.7	2.6	8.7	5.4

Table 2: Mean absolute error (averaged over the whole brain and subjects) between predicted extrapolated image and real one (values are scaled by a factor 1e3). The proposed GP model significantly outperformed predictions obtained from GLM when trained on 4 and 5 time points, from baseline to 2-3 years follow-up (* for statistically significant difference, $p < 0.05$, paired t-test).

indicates that the GP model is able to capture meaningful accelerations of the time process when sufficient data is provided, while it stays essentially linear otherwise. Figure 2 shows the mean hippocampal progression and associated confidence interval from the posterior latent process for a MCIc patient. We observe that the GP-based model of hippocampal loss is non-linear and fairly predicts the acceleration of volume loss observed in the follow-up testing images.

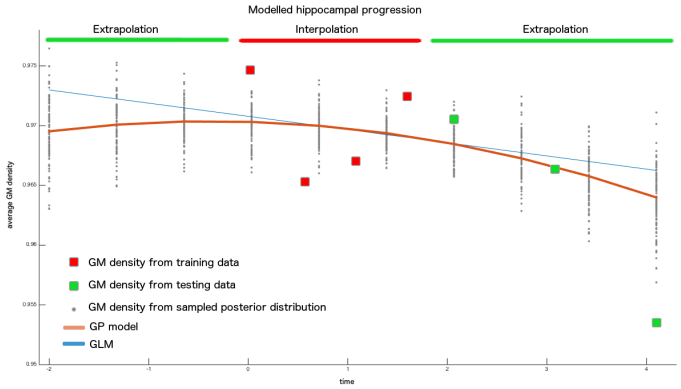


Figure 2: Predicted hippocampal progression for a sample MCIc patient. The model was estimated from 4 image time points (baseline to 2 years) in a bounding region including the hippocampus. The longitudinal sample distribution (gray dots) and mean prediction (red line) are estimated according to the marginal GP posterior of Section 3 by using the Hoffman-Ribak method.

5 Application: Between-Subjects Prediction of Individual Rates of Ventricle Growth using Multi-Kernel Learning

In this second application, we exploit the flexibility of our model to make covariate-based predictions of individual rates of atrophy in elderly subjects. In contrast to typical multivariate models which predict or classify scalar values, our GP framework allows prediction of images. In particular, we here focus on predicting the rate of volumetric growth in the lateral ventricle regions.

Firstly, we used computational morphometry to obtain the rates of atrophy in a large sample from the ADNI longitudinal dataset. To obtain these features for training and testing, we used 1143 and 569 MRI scans of 206 and 105 elderly subjects respectively (ages 59-91, age mean $\pm$ std: 76.0 $\pm$ 6.0 years). In order to enable predictions across a broad range of clinical states, the sample was pooled across clinical groups. It contained 111 healthy elderly and 108 subjects with stable and 92 subjects with progressive MCI. After longitudinal registration, tissue segmentation and inter-subject alignment [12], we calculated each subject's ventricle growth rate from registered CSF images using a linear model.

Secondly, using the preprocessed images as features we considered a special case of generative model (1) to implement a prediction model based on individual subject's covariates, e.g. age, cognitive scores, etc. This is realized by a different choice of covariance function Σ_T compared to the above within-subject application. In order to enable a prediction based on multiple available

covariate sets e.g. genes, clinical scores, etc. we used an additive multi-kernel learning covariance

$$\Sigma_T = \sum_{r=1}^{4} K_r, \quad \text{with} \quad K_r(c_1, c_2) = \alpha_r \exp(-\frac{1}{2}(c_1 - c_2)^T M_r (c_1 - c_2)) \tag{7}$$

using a sum of (up to four) squared exponential covariances K_r with amplitudes α_r, and c_1, c_2 denoting pairs of covariate vectors from each of (up to four) covariate sets. The symmetric matrices M_r were chosen to be either $M_{ISO} = \ell^{-2} Id$ or $M_{ARD} = \text{diag}(\boldsymbol{\ell})^{-2}$. Like in typical GP regression applications, using (7) explicitly models covariance of (latent) observations f as a function of similarity of inputs c (here the covariate vectors of subjects). That implements the idea that subjects with similar covariates are expected to have similar rates of atrophy. In particular, the choice of $M_r = M_{ISO}$ parametrizes an isotropic covariance assuming equal length-scale for different covariates of the same covariate set. An alternative choice of $M_r = M_{ARD}$ implements automatic relevance determination (ARD) with separate length-scales estimated for each variable. We compared successively complex prediction models using (1) only global brain volumes (tgmv, twmc, tcsv) or (2) additionally using demography (age, sex, education, marital status, year of retirement), or (3) also including genetic risk in terms of the number of ApoE4 allele and (4) finally also using the clinical neuropsychological test scores MMSE, ADAS, and CDR. The models (1) to (4) step-by-step increased the amount of subject-specific information to predict maps of rates of ventricle growth. Comparison across models was performed using log marginal likelihood balancing model fit and model complexity with varying numbers of hyperparameters. We found an increasing marginal likelihood for more complex models using ARD covariance (see Table 3) and decreased model evidence for model 4 under ISO covariance. Highest marginal likelihood was observed for ARD model 4 including all predictors. This trend is also reflected in terms of mean absolute error maps demonstrating increased prediction accuracy and generalization ability during testing in an independent test sample of 105 subjects (Figure 3A). Results also showed a correlation of up to 0.52 of predicted and true growth rates (Figure 3B).

6 Conclusions

We presented a novel framework for modelling and prediction of spatio-temporal processes in image time series. It is flexible and computationally efficient thanks to the proposed Kronecker structure of the covariance, and to the use of the Hoffman-Ribak method for efficient sampling from the posterior. Our model provided promising results when tested in very different experimental

model	ml - ISO	ml - ARD	mae - ARD
1	1.6697	1.6769	0.0059
2	2.4309	2.0249	0.0058
3	2.4356	2.0513	0.0080
4	2.2768	2.4434	0.0057

Table 3: Log marginal likelihood (ml) of Gaussian process covariance using M_{ISO} and M_{ARD} for prediction of ventricle growth rate maps based on sets of subject's covariates. Hyperparameters were optimized in 206 subjects training sample. Column 3 shows mean absolute error (mae) averaged across voxels in prediction of unseen 105 test subjects from independent test sample.

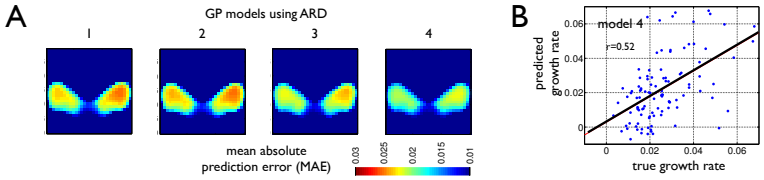


Figure 3: (A) Mean absolute error (MAE) of prediction maps in an independent testing sample of 105 subjects show increasingly better predictions using more predictor sets and Gaussian process models with ARD. (B) Predicted over true growth rates using model 4 in an example voxel showing correlation of $r = 0.52$.

scenarios concerning longitudinal modelling in AD, and opens the path to the effective use of GPs for the generative modeling of neuroimaging data. The strength of the framework relies on assuming separability of spatial and temporal processes. We show that this assumption leads to meaningful results when applied to the longitudinal modeling in AD, where the expected pathological changes are generally mild and slowly varying across brain regions. This assumption might be relaxed in future work in order to also model spatially varying processes that might underlie biological progressions with different properties. It may be indeed possible to further extend the framework to allow non-stationary correlations and noise models without compromising the computational efficiency, by accounting for local smoothly varying stationary processes as previously proposed in geostatistics [13]. Finally, further extensions of the proposed work will be devoted to the group-wise non-parametric mixed-effect modeling of disease progression in clinical cohorts such as ADNI, by exploiting the flexibility of the proposed spatio-temporal covariance structure in accounting for subject and group-specific progressions and confounders.

6.0.1 Acknowledgements

Marco Lorenzi is grateful to Prof. John Ashburner, for his help in finalizing this work, and to Dr. Richard Turner, for his precious suggestions on the train toward London. Sebastien Ourselin receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative- BW.mn.BRC10269). Gabriel Ziegler is supported in part by the German Academic Exchange Service (DAAD). The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust [grant number 091593/Z/10/Z].

References

[1] Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S.C.: Population shape regression from random design data. IJCV **90**(2) (2010) 255–266

[2] Ashburner, J., Ridgway, G.: Symmetric diffeomorphic modeling of longitudinal structural MRI. Frontiers in Neuroscience **6**(197) (02 2013)

[3] Niethammer, M., Huang, Y., Vialard, F.X.: Geodesic regression for image time-series. In: MICCAI. (2011) 655–662

[4] Lorenzi, M., Ayache, N., Frisoni, G.B., Pennec, X.: Mapping the effects of Aβ 1- 42 levels on the longitudinal changes in healthy aging: hierarchical modeling based on stationary velocity fields. In: MICCAI. (2011) 663–670

[5] Friston, K.J., Holmes, A., Worsley, K.J.: Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping **2** (1995) 189–210

[6] Flandin, G., Penny, W.D.: Bayesian fMRI data analysis with sparse spatial basis function priors. NeuroImage **34**(3) (2007) 1108–1125

[7] Harrison, L.M., Green, G.G.: A Bayesian spatiotemporal model for very large data sets. NeuroImage **50**(3) (2010) 1126–1141

[8] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2005)

[9] Ziegler, G., Ridgway, G.R., Dahnke, R., Gaser, C.: Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. NeuroImage **97** (April 2014) 333–348

[10] Stegle, O., Lippert, C., Mooij, J.M., et al.: Efficient inference in matrix-variate gaussian models with iid observation noise. In: Advances in Neural Information Processing Systems 24

[11] Hoffman, Y., Ribak, E.: Constrained realizations of Gaussian fields - A simple algorithm. Astrophys. J. Lett. **380** (October 1991) L5–L8

[12] Ashburner, J., Friston, K.: Unified segmentation. NeuroImage **26** (2005) 839–851

[13] Gelfand, A., Fuentes, M., Guttorp, P., Diggle, P.: Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis (2010)

Selected publications for Axis II

Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data

Luigi Antelmi[1], Nicholas Ayache[1], Philippe Robert[2] and Marco Lorenzi[1]
1. Université Côte d'Azur, INRIA Sophia Antipolis, EPIONE research group, France
2. Centre Memoire, CHU of Nice, France

Originally published in:
Proceedings of the 36th International Conference on Machine Learning, ICML, 2019

Abstract

Interpretable modeling of heterogeneous data channels is essential in medical applications, for example when jointly analyzing clinical scores and medical images. Variational Autoencoders (VAE) are powerful generative models that learn representations of complex data. The flexibility of VAE may come at the expense of lack of interpretability in describing the joint relationship between heterogeneous data. To tackle this problem, in this work we extend the variational framework of VAE to bring parsimony and interpretability when jointly account for latent relationships across multiple channels. In the latent space, this is achieved by constraining the variational distribution of each channel to a common target prior. Parsimonious latent representations are enforced by variational dropout. Experiments on synthetic data show that our model correctly identifies the prescribed latent dimensions and data relationships across multiple testing scenarios. When applied to imaging and clinical data, our method allows to identify the joint effect of age and pathology in describing clinical condition in a large scale clinical cohort.

1 Introduction

Understanding the relationship among heterogeneous data is essential in medical applications, where performing a diagnosis, or understanding the dynamics of a pathology require to jointly analyze multiple data channels, such as demographic data, medical imaging data, and psychological tests.

Multivariate methods to jointly analyze heterogeneous data, such as Partial Least Squares (PLS), Reduced Rank Regression (RRR), or Canonical correlation analysis (CCA) [8] have successfully been applied in biomedical research [18], along with multi-channel [11, 19] and non-linear [9, 2] variants. They are classified as *recognition* methods, as their common formulation consists in projecting the observations in a latent low dimensional space in which desired characteristics are enforced, such as maximum correlation (CCA), maximum covariance (PLS), or minimum regression error (RRR) [7]. These models, though, are not *generative* because they do not provide an explicit formulation to sample observations when the distribution of latent variables and parameters is known. *Bayesian-CCA* [17] actually goes in this direction: it is a generative formulation of CCA, where a transformation of a latent variable captures the shared variation between data channels.

A limitation of this method for the application in real data scenarios is scalability, as inference on the posterior distribution results in $\mathcal{O}(D^3)$ complexity, being D the dimensionality of the data. Consequently, all the practical applications of Bayesian CCA in the earlier works were limited to very few dimensions and channels [16].

Variational autoencoders (VAEs) [15, 22] are models that couple a recognition function, or *encoder*, to infer a lower dimensional representation of the data, with a generative function, or *decoder*, which transform the latent representation back to the original observation space. The VAE is a Bayesian model: the latent variables are inferred by estimating the associated posterior distributions. Inference is efficiently performed through *amortized inference* [12] by parametrizing the posterior moments with neural networks. The networks are variationally optimized to maximize the associated evidence lower bound (ELBO). VAEs are flexible and can account for any kind of data. Within this setting, the joint analysis of heterogeneous channels can be performed through concatenation of the different data sources. However, modeling concatenated multi-channel data through a VAE may pose interpretability issues, as it is difficult to disentangle the contribution of a single channel in the description of the latent representation. Moreover, at test time, the model can only be applied to data presenting all the channels information.

To tackle this problem, in this work we generalize the VAE by assuming that in a multi-channel scenario the latent representation associated to each channel must match a common target distribution. This is done by imposing a constraint on the latent representations in an information theoretical sense, where each latent representation is enforced to match a common target prior. We will show that this constraint can be optimized within a variational optimization framework, allowing efficient inference of channel encodings and latent representation.

Another limitation of the VAE concerns the interpretability of the latent space. In particular, we generally lack of a theoretical justification for the choice of the dimensionality of the latent space. This is a key parameter that can profoundly impact the interpretability of the estimated data representation. The optimization of this latent dimension through cross-validation may also pose generalization problems, especially when the data is scarce. To tackle this issue, in this work we investigate a principled theoretical framework for imposing parsimonious representations of the latent space through sparsity. We argue that such a constraint may result not only in improved interpretability, but also in optimal data representation. Indeed, it is known that VAEs suffers from the problem of *over-pruning*: the variational approximation leads to overly simplified representations, resulting in high model bias due to the impossibility to learn latent distribution different from the prior [5, 1]. As discussed in [27], over-pruning is a recurrent phenomenon ultimately leading to excessive regularization, even in cases when the model underfits the data. The authors tackle over-pruning with the introduction of a categorical sampler on the latent space dimensions. Another way to tackle over-pruning is to enforce sparsity on the latent space. Recently [14, 21] showed that *dropout*, a technique that regularize neural networks, can be naturally encoded in VAE to lead to a sparse representation of the variational parameters.

In our work, we leverage on these recent results to enforce sparsity on the proposed multi-channel VAE. In the variational formulation, the dropout parameters are not hyperparameters anymore, and can be directly learned through the optimization of the variational constraint.

The rest of the paper is organized as follows. In Section 2 we first describe the Multi-Channel Variational Autoencoder and mathematically derive the variational constraint as an extension of the VAE framework. The sparse representation of the latent space is further analysed and discussed. In Section 3 we show results on extensive synthetic experiments comparing our model to standard non-sparse VAE formulations. We conclude the Section with the application of our model

to real data, related to clinical cases of brain neurodegeneration. We show how the learned dropout parameter can be used to automatically identify meaningful latent effect of age and pathology, allowing to predict clinical diagnosis in Alzheimer's Disease (AD). Finally, we summarize our work and propose future extensions.

2 Method

We first describe the proposed Multi-Channel Variational Autoencoder (§2.1) In §2.2 we show how out mothod can be enforced to be sparse.

2.1 Multi-Channel Variational Autoencoder

Let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_C\}$ be an observation set of C channels, where each $\mathbf{x}_c$ is a d-dimensional vector. Also, let $\mathbf{z}$ denote the l-dimensional latent variable commonly shared by each $\mathbf{x}_c$. We assume the following generative process for the observation set:

$$\begin{aligned} &\mathbf{z} \sim p(\mathbf{z}) \\ &\mathbf{x}_c \sim p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c) \qquad \text{for } c \text{ in } 1 \ldots C \end{aligned} \tag{1}$$

where $p(\mathbf{z})$ is a prior distribution for the latent variable and $p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c)$ is a likelihood distribution for the observations conditioned on the latent variable. We assume that the likelihood functions belong to a distribution family $\mathcal{P}$ parametrized by the set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}$.

In the scenario depicted so far, solving the inference problem allows the discovery of the common latent space from which the observed data in each channel is generated. The solution to the inference problem is given by deriving the posterior $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, that is not always computable analytically. In this case, *Variational Inference* can be applied to compute an approximate posterior [4].

Our working hypothesis is that every channel brings by itself some information about the latent variable distribution. As such, it make sense to approximate the posterior distribution with $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$, by conditioning it on the single channel $\mathbf{x}_c$ and its variational parameter $\boldsymbol{\phi}_c$. Since each channel provides a different approximation, we can impose a constraint enforcing each $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ to be as close as possible to the target posterior distribution. Being the mismatch measured in terms of Kullback-Leibler (KL) divergence, we specify this constraint as:

$$\underset{q \in \mathcal{Q}}{arg\,min}\ \mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||\, p(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_C, \boldsymbol{\theta})\big) \right] \tag{2}$$

where the approximate posteriors $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ belong to a distribution family $\mathcal{Q}$ parametrized by the set of parameters $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C\}$, and represent the view on the latent space that can be inferred from each channel $\mathbf{x}_c$. Practically, solving the objective in Eq. (2) allows to minimize the discrepancy between the variational approximations and the target posterior. In §2.1.1 we show that the optimization (2) is equivalent to the optimization of the following evidence lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = \mathbb{E}_c \left[L_c - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||\, p(\mathbf{z})\big) \right], \tag{3}$$

where $L_c = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)} \sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\theta}_c)$ is the expected log-likelihood of decoding each channel from the latent representation of the channel $\mathbf{x}_c$ only. This formulation is valid for any distribution family $\mathcal{P}$ and $\mathcal{Q}$.

2.1.1 Derivation of the Evidence Lower Bound

In the following derivation we will omit the variational and generative parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ to leave the notation uncluttered.

The formula in (2) states that variational inference is carried out by introducing a set of probability density functions $q(\mathbf{z}|\mathbf{x}_c)$, belonging to a distribution family $\mathcal{Q}$, that are as close as possible to the target posterior over the latent variable $p(\mathbf{z}|\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_C\})$. Given the intractability of $p(\mathbf{z}|\mathbf{x})$ for most complex models, we cannot solve directly this optimization problem. We look then for an equivalent problem, by rearranging the objective:

$$
\begin{aligned}
&\mathbb{E}_c\left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c)\,||\,p(\mathbf{z}|\mathbf{x})\big)\right] = \\
&= \mathbb{E}_c \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}_c)\left(\ln q(\mathbf{z}|\mathbf{x}_c) - \ln p(\mathbf{z}|\mathbf{x})\right) d\mathbf{z} \\
&= \mathbb{E}_c \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}_c) \\
&\qquad \left(\ln q(\mathbf{z}|x_c) - \ln p(\mathbf{x}|\mathbf{z}) - \ln p(\mathbf{z}) + \ln p(\mathbf{x})\right) d\mathbf{z} \\
&= \ln p(\mathbf{x}) + \\
&\qquad \mathbb{E}_c\left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c)\,||\,p(\mathbf{z})\big) - \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)}\left[\ln p(\mathbf{x}|\mathbf{z})\right]\right]
\end{aligned}
$$

where we factorize the true posterior $p(\mathbf{z}|\mathbf{x})$ using the Bayes' theorem. Now, we can reorganize the terms, such that:

$$
\begin{aligned}
&\ln p(\mathbf{x}) - \underbrace{\mathbb{E}_c\left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c)\,||\,p(\mathbf{z}|\mathbf{x})\big)\right]}_{\geq 0} = \\
&= \underbrace{\mathbb{E}_c\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)}\left[\ln p(\mathbf{x}|\mathbf{z})\right] - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c)\,||\,p(\mathbf{z})\big)\right]}_{\text{lower bound } \mathcal{L}}
\end{aligned}
\tag{4}
$$

Since the KL term in the left hand side is always non-negative, the right hand side is a lower bound to the log evidence. Thus, by maximizing the lower bound we also maximize the data log evidence while solving the minimization problem in (2).

We note that the lower bound (4) is composed by a regularization term and a data matching term. The $\mathcal{D}_{KL}$ term minimizing the mismatch between the approximate distribution and the target prior, acts as a regularizer. The inner expectation term favors the approximate posterior that maximize the data log-likelihood.

The hypothesis that every channel is conditionally independent from all the others given $\mathbf{z}$, allows to factorize the data likelihood as $p(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{C} p(\mathbf{x}_i|\mathbf{z})$, so that the lower bound becomes:

$$
\begin{aligned}
&\mathcal{L} = \mathbb{E}_c\left[L_c - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c)\,||\,p(\mathbf{z})\big)\right] \\
&\text{where} \quad L_c = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)}\left[\sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z})\right].
\end{aligned}
$$

2.1.2 Comparison with VAE

Our model extends the VAE [15]: the novelty is in the log-likelihood terms L_c in Eq. (3), representing the reconstruction of the whole multi-channel dataset from a single channel only. In case $C = 1$ the

model collapses to a VAE. In the case $C > 1$, the L_c terms considered altogether force each channel to the joint decoding of itself and every other channel at the same time. This characteristic allows, on a testing dataset, to reconstruct missing channels. Hence, our model is different from a VAE where all the channels have been concatenated into a single one. In the latter naive application of the VAE there cannot be missing channels if we want to infer the latent space variables. Our model is also different from a stack of C independent VAEs, in which the C latent spaces are no more related to each-other. The dependence between encoding and decoding across channels stems from the joint approximation of the posterior distribution (Formula (2)).

2.1.3 Gaussian linear case

Model (1) is completely general and can account for complex non-linear relationships modeled, for example, through deep neural networks. However, for simplicity of interpretation, in what follows we focus our multi-channel variational framework to the *Gaussian Linear Model*. This is a special case, analogous to Bayesian-CCA [17], where the members of the variational family $\mathcal{Q}$ and generative family $\mathcal{P}$ are Gaussian parametrized by linear transformations. We define the members of the families $\mathcal{Q}$ and $\mathcal{P}$ as:

$$q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) = \mathcal{N}\left(\mathbf{z}|\mathbf{V}_c^{(\mu)}\mathbf{x}_c, diag(\mathbf{V}_c^{(\sigma)}\mathbf{x}_c)\right) \tag{5}$$

$$p\left(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c\right) = \mathcal{N}\left(\mathbf{x}_c|\mathbf{G}_c^{(\mu)}\mathbf{z}, diag(\mathbf{g}_c^{(\sigma)})\right) \tag{6}$$

i.e. factorized multivariate Gaussian distributions whose first moments are linear transformations depending on the conditioning variables. $\boldsymbol{\theta}_c = \{\mathbf{G}_c^{(\mu)}, \mathbf{g}_c^{(\sigma)}\}$ and $\boldsymbol{\phi}_c = \{\mathbf{V}_c^{(\mu)}, \mathbf{V}_c^{(\sigma)}\}$ are the parameters to be optimized by maximizing the lower bound in (3).

2.1.4 Optimization of the lower bound

The optimization starts with a random initialization of the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}$ and $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C\}$. The expectations L_c in the Eq. (3) can be computed by sampling from the variational distributions $q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right)$ and, when the prior $p\left(\mathbf{z}\right) = \mathcal{N}\left(\mathbf{0}; \mathbf{I}\right)$, the $\mathcal{D}_{KL}$ term in Eq. (3) can be computed analytically (*cf.* [15], appendix 2.A). The maximization of $\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}\right)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is efficiently carried out through minibatch stochastic gradient descent implemented with the backpropagation algorithm. With *Adam* [13] we compute adaptive learning rates for the parameters.

2.2 Inducing Sparse Latent Representations

In extensive synthetic experiments with the non-sparse version of the multi-channel model, we found that generally the lower bound reaches the maximum value at convergence when the number of fitted latent dimensions coincide with the true one used to generate the data (see *Sup. Mat.*). This procedure provides an heuristic for selecting the latent variable dimensions, and proved to work well in controlled scenarios. However, according to our experience, it fails in most complex cases (see *Sup. Mat.*), and is time consuming. Moreover, our trust in the result depends on the tightness between the model evidence and its lower bound: a factor that is not easy to control. To address this issue, we propose here to automatically infer the latent variable dimensions via a sparsity constraint on $\mathbf{z}$. Moreover, having a sparse $\mathbf{z}$ as a direct result of one single optimization would

be computationally advantageous. Likewise, it would ease the interpretability of the observation model in (1) as the number of relationships to take into account decreases.

2.2.1 Regularization via Dropout

Dropout [23] and *DropConnect* [25] are techniques for regularizing neural networks. The basic block of a neural network is the *fully connected* layer, composed by a linear transformation of an input vector $\mathbf{a}$ into an output vector $\mathbf{b}$, and a non linearity applied to the components of $\mathbf{b}$. Given a generic linear transformation $\mathbf{b} = \mathbf{W}\mathbf{a}$, with $\mathbf{a}$ and $\mathbf{b}$ column vectors, regularization techniques are based on the multiplication of either $\mathbf{a}$ (dropout) or $\mathbf{W}$ (dropconnect) by independent Bernoulli random variables. The components of $\mathbf{b}$ are hence computed as:

$$b_i = \sum_k w_{ik}(\xi_k a_k) \qquad \text{(dropout)} \tag{7}$$

$$b_i = \sum_k (\xi_{ik} w_{ik}) a_k \qquad \text{(dropconnect)} \tag{8}$$

where $\xi_k, \xi_{ik} \sim \mathcal{B}(1-p)$ with hyperparameter p known as *drop rate*. The elements b_i are approximately Gaussian for the Lyapunov's central limit theorem [26], and their distributions takes the form:

$$b_i \sim \mathcal{N}\left(\textstyle\sum_k \theta_{ik}; \alpha \sum_k \theta_{ik}^2\right) \tag{9}$$

where $\alpha = {}^{p}/_{1-p}$ and $\theta_{ik} = w_{ik} a_k (1-p)$. In *Gaussian dropout* (*ibid.*) the regularization is achieved by sampling directly from (9). It's easy to verify that if we sample the components of $\mathbf{a}$ from a Gaussian with variance proportional to the square of their expected values, i.e. if:

$$a_i \sim \mathcal{N}\left(\mu_i; \alpha \mu_i^2\right) \tag{10}$$

expression (9) still holds with $\theta_{ik} = w_{ik}\mu_k$, keeping unchanged the connection with the original dropout techniques.

2.2.2 Variational Dropout and Sparsity

In the context of the *Variational Autoencoder* (VAE), posterior distributions that takes the form as in (10) are called *dropout posteriors* [14]. More precisely, if the variational posteriors on the network weights w are dropout posteriors, Gaussian dropout arises from the application of the *local reparameterization trick*, a method introduced to increase the efficiency of gradients estimation in training. The only prior on w consistent with the optimization of the lower bound is the improper log-scale uniform:

$$p\left(\ln|w|\right) = \text{const} \Leftrightarrow p\left(|w|\right) \propto \frac{1}{|w|} \tag{11}$$

With this prior, the KL divergence of the dropout posterior depends only on α and can be approximated numerically. In [21] the authors provide an approximation of the KL (reported in (12) to allow this parameter to be learned through the optimization of the lower bound via

gradient-based methods:

$$\begin{gathered}\mathcal{D}_{KL}\big(\mathcal{N}\left(w;\alpha w^2\right)||\,p\left(w\right)\big)\approx\\ \approx -k_1\sigma(k_2+k_3\ln\alpha)+0.5\ln(1+\alpha^{-1})+k_1\\ k_1=0.63576\quad k_2=1.87320\quad k_3=1.48695\\ \sigma(\cdot)\text{ Sigmoid function}\end{gathered}\tag{12}$$

While the optimization of the KL divergence promotes $\alpha\to\infty$, the implicit drop rate p tends to 1, meaning that that particular weight w can be discarded. Sparsity arises naturally: large values of w correspond to even larger uncertainty αw^2 because of the quadratic relationship and the tendency of the optimization objective to favors $\alpha\to\infty$; therefore, unless that weight is beneficial for the second optimization objective, that is to maximize the data log-likelihood, it will be set to zero.

2.2.3 Sparse Multi-Channel VAE

Compatibly with standard dropout methods, in our Multi-Channel VAE we define a variational approximation of the latent code $\mathbf{z}$. We note that the local reparameterization trick cannot be straightforwardly applied, since it would require to transfer the uncertainty to a lower dimensional variable, such as from $\mathbf{W}$ to $\mathbf{b}$ in §2.2.1. We notice however that by choosing a dropout posterior for $q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)$, the weight of the decoding transformation transforming $\mathbf{z}$ to the output are Bernoulli distributed $\sim\mathcal{B}(1-p)$. Therefore, we established an analogy with the standard dropout techniques. Specifically, imposing a dropout posterior for the latent code $\mathbf{z}$ is analogous to perform dropout on the latent code itself, and dropconnect on the decoder weights. We therefore define the approximate posteriors $q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)$ in Eq. (3) and parametrize them to be factorized dropout posteriors, that is, for c in $1\dots C$:

$$q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)=\mathcal{N}\left(\boldsymbol{\mu}_c;\boldsymbol{\alpha}\odot\operatorname{diag}(\boldsymbol{\mu}_c)^2\right)\tag{13}$$

with $\boldsymbol{\mu}_c=\boldsymbol{\phi}_c\mathbf{x}_c$, where parameters $\boldsymbol{\phi}=\{\boldsymbol{\alpha},\boldsymbol{\phi}_1,\dots,\boldsymbol{\phi}_C\}$ include $\boldsymbol{\phi}_c$ linear transformations, specific to channel c, while $\boldsymbol{\alpha}$ is shared among all the channels. Following the considerations of [14], the prior distribution $p\left(\mathbf{z}\right)$ is chosen to be fully factorized by scale-invariant log-uniform priors:

$$p\left(\mathbf{z}\right)=\prod p\left(|z_i|\right)\quad\text{such that}\quad p\left(\ln|z_i|\right)\propto\text{const}\tag{14}$$

Because of these choices, the $\mathcal{D}_{KL}$ term in Eq. (3) can be easily computed by leveraging on Eq. (12). For the same considerations made in the previous section, we induce a sparse behavior on the components of $\mathbf{z}$ and on the associated decoder parameters (*cfr.* Fig. 1). The variational parameter $\boldsymbol{\alpha}$ con be learned, and as the connection with the dropout techniques is kept, we can leverage on the relationship between $\boldsymbol{\alpha}$ and the dropout rate p to interpret the relative importance of the latent dimensions.

3 Experiments

We first describe our results on extensive synthetic experiments conducted with our non sparse model and its sparse variant. We benchmark these models with the VAE and conclude the Section with the application of our sparse model to real data, related to clinical cases of neurodegeneration.

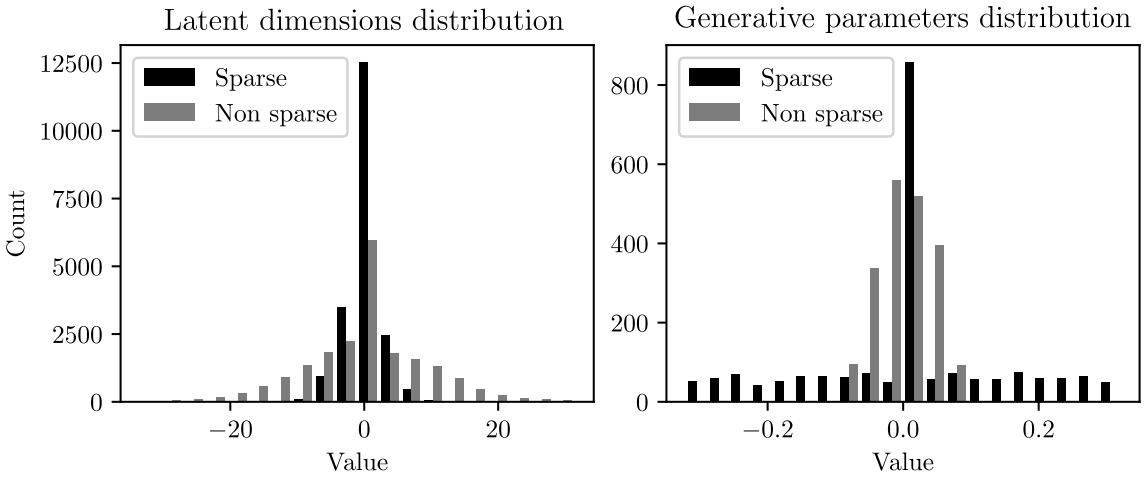


Figure 1: Effect of variational dropout on a synthetic experiment. As expected, the minimum number of non-zero components of $\mathbf{z}$ (left) and generative parameters (right) is obtained with the sparse model.

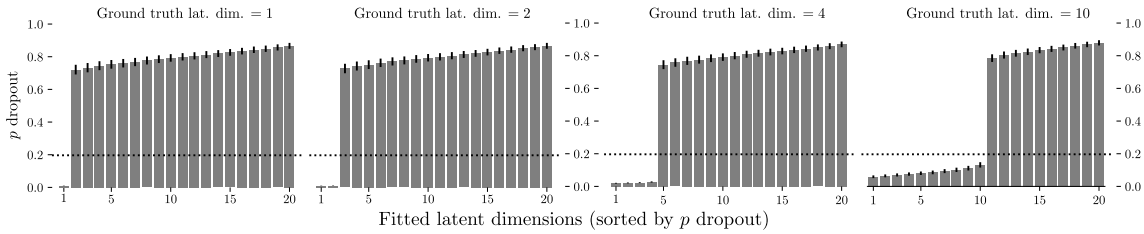


Figure 2: Estimated dropout rates for the latent dimensions when applying the Sparse Multi-Channel VAE on data generated with respectively 1, 2, 4, and 10 latent dimensions.

3.1 Synthetic Experiments

Datasets $\mathbf{x} = \{\mathbf{x}_c\}$ with $c = 1 \dots C$ channels where created according to the following model:

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}_l) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}_{d_c}) \\ \mathbf{G}_c &= diag\left(\mathbf{R}_c \mathbf{R}_c^T\right)^{-1/2} \mathbf{R}_c \\ \mathbf{x}_c &= \mathbf{G}_c \mathbf{z} + snr^{-1/2} \cdot \boldsymbol{\epsilon} \end{aligned} \tag{15}$$

where for every channel c, $\mathbf{R}_c \in \mathbb{R}^{d_c \times l}$ is a random matrix with l orthonormal columns (*i.e.,* $\mathbf{R}_c^T \mathbf{R}_c = \mathbf{I}_l$), $\mathbf{G}_c$ is the linear generative law, and snr is the signal-to-noise ratio. With this choice the diagonal elements of the covariance matrix of $\mathbf{x}_c$ are inversely proportional to snr, *i.e.,* $diag\left(\mathbb{E}\left[\mathbf{x}_c \mathbf{x}_c^T\right]\right) = (1 + snr^{-1})\mathbf{I}_{d_c}$. Scenarios where generated by varying one-at-a-time the dataset attributes, as listed in Tab. 1.

Table 1: Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (15).

Attribute description	Iteration list
Total channels (C)	2 3 5 10
Channel dimension (d_c)	32
Latent space dimension (l)	1 2 4 10 20
Number of samples/observations	100 1000
Signal-to-noise ratio (snr)	10 1
Seed (re-initialize $\mathbf{R}_c$)	1 2 3 4 5

ELBO in non-sparse Multi-Channel VAE. For each generated scenario, we optimized multiple instances of a Gaussian Linear Multi-Channel model, as defined in §2.1.3. At convergence, the loss function (negative lower bound) has a minimum when the number of fitted latent dimensions corresponds to the number of the latent dimensions used to generate the data. When increasing the number of fitted latent dimensions, a sudden decrease of the loss (*elbow* effect) is indicative that the true number of latent dimensions has been found. These results are summarized in the *Supplementary Materials*, where we show also that the elbow effect becomes more evident when increasing the number of channels. Ambiguity in identifying the elbow may arise for high-dimensional data channels.

3.2 Sparse Multi-Channel VAE Benchmark

This benchmark is based on the data scenarios illustrated in the previous section (Tab. 1). For each generated dataset, we optimized our Multi-Channel VAE with dropout posteriors (eq. 13) associated to log-uniform priors as in (eq. 14).

Results. In Fig. 1 we compare the latent space distributions and the generative parameters derived from the application of the sparse and non-sparse Multi-Channel VAE, after fitting the two

models on the same data. As expected, the number of zero elements is considerably higher in the sparse version.

In all the synthetic scenarios, the models are optimized by imposing a 20-dimensional latent space. We note that the learned dropout rate is very low for the dimensions corresponding to the true latent dimensions used to generate the fitted scenario (Fig. 2). Because of this, model selection can be performed by retaining those latent dimensions satisfying an opportune threshold on the dropout rates. We can see that with the threshold $p < 0.2$, is possible to safely recover the true number of latent dimensions across all the testing scenarios.

3.3 Comparison with VAE

We compared the performance of four variational methods applied to the synthetic scenarios. Besides our sparse and non-sparse Multi-Channel models, we considered a VAE, and a stack of independent VAEs (IVAEs). In the VAE cases, channels where concatenated feature-wise to form a single channel. In IVAEs experiments, every channel was independently modeled with a VAE. The comparison metric is the *mean absolute error* between the generated testing data and then predictions from the inferred latent space.

Results. As depicted in Fig. 3, sparse Multi-Channel models performs consistently better than the non-sparse ones. In high *snr* cases, the sparse model performs equivalently or better than the VAE. The IVAEs models leads to the worst performances. This is expected, as the ground truth data variability depends on the joint information across channels. By modeling each channel independently, part of this variability is therefore mistaken as noise.

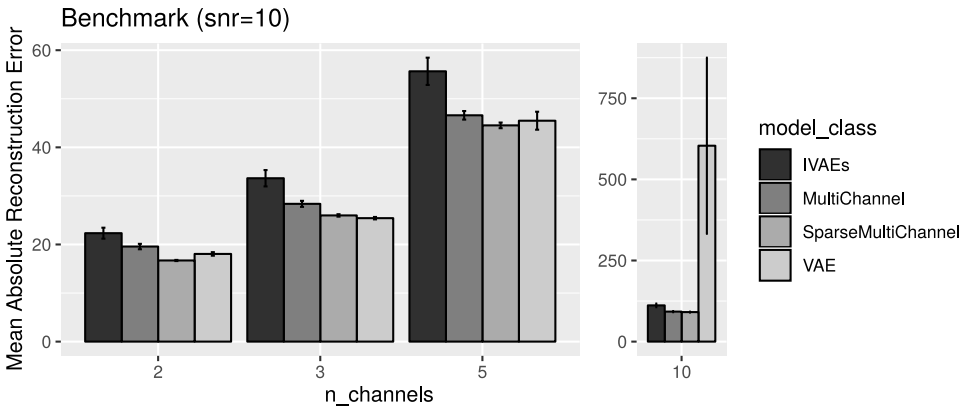
3.4 Medical Imaging data

3.4.1 Data preparation

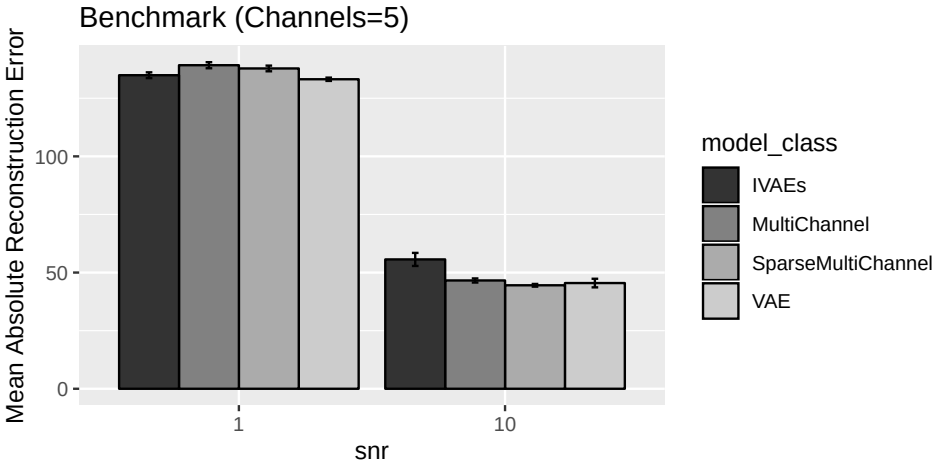
Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.

We analyzed clinical and imaging channels from 504 subjects of the ADNI cohort. We randomly assigned the subjects to a training and testing set through 10-fold cross validation. The clinical channel was composed by six continuous variables generally recorded in memory clinics: age; results to mini-mental state examination, adas-cog, cdr, and faq tests; scholarity level. The three imaging channels were structural MRI (gray matter only), functional FDG-PET, and Amyloid-PET. Raw data from the imaging channels were coregistered in a common geometric space by means of voxel-based morphometry methods [3]. Visual quality check was performed to exclude processing errors. Image intensities were finally averaged over 90 brain regions mapped in the AAL atlas [24] to produce 90 features arrays for each image. Lastly, data was centered and standardized across features. Our sparse multi-channel model (§2.2.3) was optimized on the resulting multi-channel dataset.

Results. We identify 5 optimal latent dimensions, by applying the dropout threshold of 0.2 as identified in the synthetic experiments (Fig. 2).



(a)



(b)

Figure 3: Benchmark of four variational methods applied to the multi-channel scenarios in Tab. 1. Sparse Multi-Channel models performs consistently better than the non-sparse ones. In high *snr* cases, the sparse model performs equivalently or better than the VAE.

The encoding of the test set in the latent space given by our model is depicted in Fig. 4, where we limited the visualization to the 2D subspace generated by the two most relevant dimensions. This subspace appears stratified by age and disease status, across roughly orthogonal directions. We note however that the model was agnostic to the disease status, and was able to correctly stratify the testing data only thanks to the learned latent representation. This is shown in Table 2, where the latent representation provided by our sparse Multi-Channel framework leads to competitive predictive performances in predicting the clinical status. Prediction was performed on the testing set via Linear Discriminant Analysis fitted on the training latent space. We note that the predictive accuracy is particularly high in the challenging case of detecting MCI and Normal groups, notoriously characterized by high heterogeneity [20].

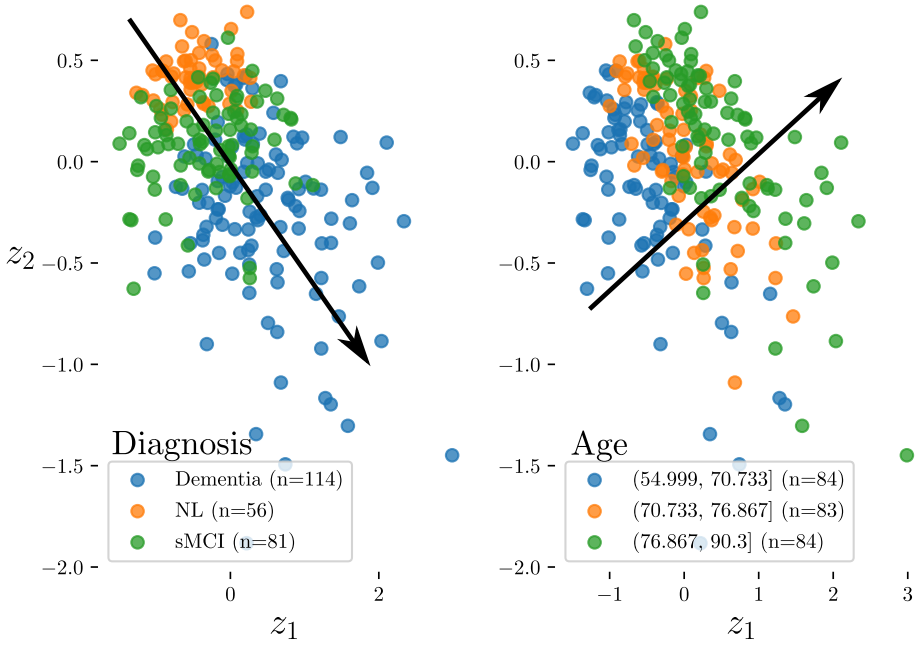
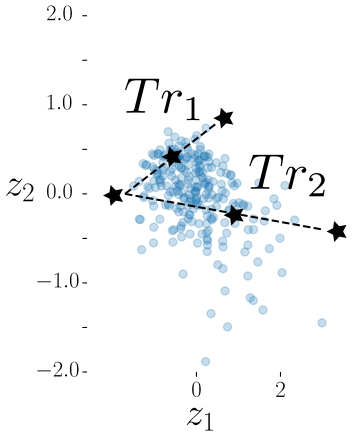
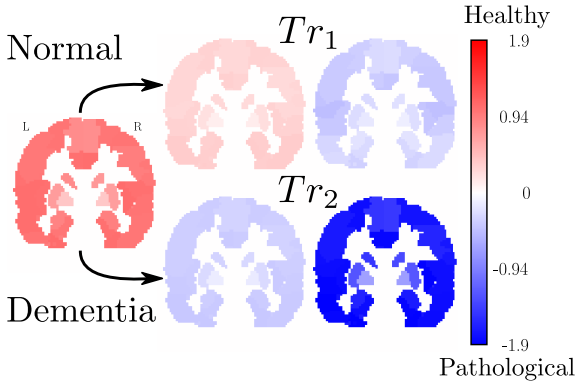


Figure 4: Stratification of the ADNI subjects (test data) in the sparse latent subspace inferred from the first two least dropped out dimensions. In the same subspace it is possible to stratify subjects in the test-set by disease status (left) and by age (right) in almost orthogonal directions. Classification accuracy for these subjects is given in the rightmost column of Tab. 2.

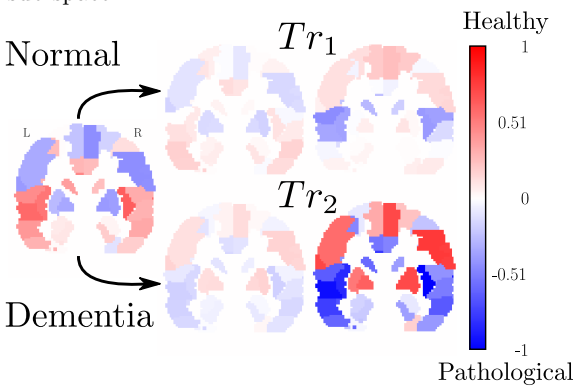
We tested the generative capability of our model, by sampling points from two trajectories in the subspace shown in Fig. 4 to predict the imaging data channels. Trajectory 1 (Tr_1) follows an aging path centered on the healthy subject group. Trajectory 2 (Tr_2), starts from the same origin of Tr_1 and follows a path were aging is entangled with the pathological variability. We can see these trajectories and the generated imaging channels in Fig. 5. Fig. 6 shows the generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (6)) of the four channels associated to the most relevant latent dimension identified by dropout. These generative parameters show a plausible relationship across channels, describing a pattern of early onset AD, associated with abnormal scores (low MMSE, high ADAS and CDR), gray matter atrophy emerging from the MRI, low glucose uptake in the temporal lobes



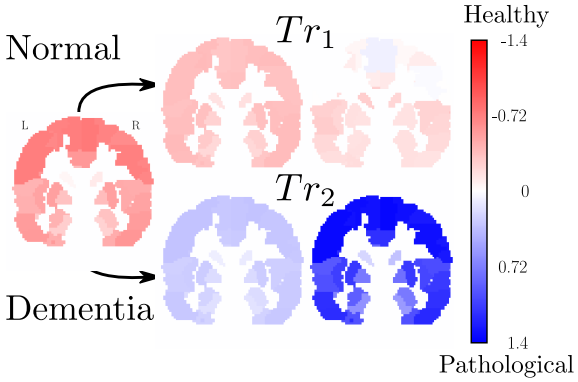
(a) Latent sub-space



(b) MRI



(c) FDG



(d) Amyloid

Figure 5: Generation of imaging data from trajectories in the latent space. (a) Normal aging trajectory (Tr_1) *vs* Dementia aging trajectory (Tr_2) in the latent 2D sub-space (*cfr.* Fig. 4). Stars indicates the sampling points along trajectories. The trajectories share the same origin. MRIs (b), FDG (c), and Amyloid PET (d). All the trajectories show a plausible evolution across disease and healthy conditions.

as emerging from the FDG-PET, and high amyloid deposits, coherently with the research literature on Alzheimer's Disease [6, 10].

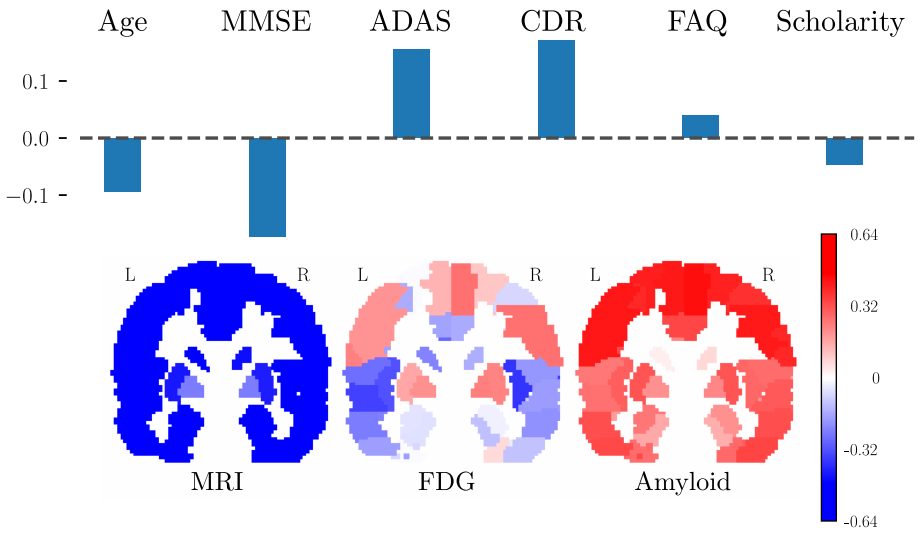


Figure 6: Generative parameters (*cfr.* $\mathbf{G}_c^{(\mu)}$ in Eq. (6)) of the four channels associated to the least dropout latent dimension in the sparse multi-channel model. (Top) Clinical channel parameters. (Bottom) Imaging channels parameters.

Table 2: Proportion of correctly classified ADNI subjects belonging to the testing hold-out dataset. Classification done by means of *Linear Discriminant Analysis* using as training data the latent space inferred with the sparse and non sparse models. 10-fold cross validation results shown as "mean (standard deviation)". Within the sparse framework, we selected the subspace generated by the most relevant latent dimensions identified by variational dropout ($p < 0.2$).

Model:	Non sparse	Sparse
z dimensions: **used** (fitted)	**16** (16)	**5** (16)
Dementia	**0.88** (**0.08**)	0.85 (0.07)
Mild Cognitive Impairment	0.58 (0.07)	**0.70** (**0.05**)
Normal	0.82 (0.07)	**0.89** (**0.03**)

4 Conclusion

This paper introduces the Sparse Multi-Channel VAE, an extension of variational autoencoders, to jointly account for latent relationships across heterogeneous data. Parsimonious and interpretable representations are enforced by variational dropout, to show how sparsity provides an effective mean to model selection in the latent space. In extensive synthetic experiments, we compared the

performance of our model against the VAE, where multi-channel data were stacked feature-wise to produce a single channel, and against an independent stack of VAEs, where each channel was treated independently. We found a comparable and most of the times superior performance of our model with respect to the benchmark. In the real case scenario of Alzheimer's Disease modeling, our model allowed the unsupervised stratification of the latent space by disease status and age, providing evidence for a clinically sound interpretation of the latent space. Future extension of this work will focus on the extension to non-linear parameterization of the variational distributions. Given the scalability of our variational model, application to high resolution images may be also at reach, although this may require to account for full covariance matrices to take into account spatial relationships. To increase the model classification performance, supervised clustering of the latent space can be introduced, for example, by leveraging on the categorical sampler introduced in the latent space by the *epitomic* VAE [27]. Lastly, due to the general formulation, the proposed method can find various applications as a general data interpretation technique, not limited to the biomedical research area.

References

[1] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. nov 2017.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep Canonical Correlation Analysis. *Proc. Mach. Learn. Res.*, 28(3):1247–1255, 2013.

[3] John Ashburner and Karl J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6 Pt 1):805–21, jun 2000.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. 2016.

[5] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. sep 2015.

[6] Bruno Dubois, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, Stefano Cappa, Sebastian Crutch, Sebastiaan Engelborghs, Giovanni B Frisoni, Nick C Fox, Douglas Galasko, Marie-odile Habert, Gregory A Jicha, Agneta Nordberg, Florence Pasquier, Gil Rabinovici, Philippe Robert, Christopher Rowe, Stephen Salloway, Marie Sarazin, Stéphane Epelbaum, Leonardo C de Souza, Bruno Vellas, Pieter J Visser, Lon Schneider, Yaakov Stern, Philip Scheltens, and Jeffrey L Cummings. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet. Neurol.*, 13(6):614–29, jun 2014.

[7] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[8] Harold Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321, dec 1936.

[9] Su-Yun Huang, Mei-Hsien Lee, and Chuhsing Kate Hsiao. Nonlinear measures of association with kernel canonical correlation analysis and applications. *J. Stat. Plan. Inference*, 139(7):2162–2174, 2009.

[10] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's Dement.*, 14(4):535–562, 2018.

[11] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[12] Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. Semi-Amortized Variational Autoencoders. feb 2018.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6, 2014.

[14] Diederik P Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett, editors, *Adv. Neural Inf. Process. Syst. 28*, pages 2575–2583. Curran Associates, Inc., 2015.

[15] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014).*, dec 2014.

[16] Arto Klami and Samuel Kaski. Local dependent components. In Zoubin Ghahramani, editor, *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, pages 425–432. Omnipress, 2007.

[17] Arto Klami, Virtanen Seppo, and Samuel Kaski. Bayesian Canonical Correlation Analysis. *J. Mach. Learn. Res.*, 14:965–1003, 2013.

[18] Jingyu Liu and Vince D. Calhoun. A review of multivariate analyses in imaging genetics. *Front. Neuroinform.*, 8:29, mar 2014.

[19] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction. *IEEE Trans. Knowl. Data Eng.*, 27(11):3111–3124, 2015.

[20] Alex F. Mendelson, Maria A. Zuluaga, Marco Lorenzi, Brian F. Hutton, and Sébastien Ourselin. Selection bias in the reported performances of AD classification pipelines. *NeuroImage Clin.*, 14:400–416, 2017.

[21] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational Dropout Sparsifies Deep Neural Networks. *arXiv*, 2017.

[22] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. jan 2014.

[23] Nitish Srivastava, Hinton Geoffrey, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014.

[24] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *Neuroimage*, 15(1):273–289, jan 2002.

[25] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of Neural Networks using DropConnect. In *Proc. 30th Int. Conf. Mach. Learn.*, pages 1058—-1066, 2013.

[26] Sida Wang and Christopher Manning. Fast dropout training. *Proc. 30th Int. Conf. Mach. Learn.*, 28(2):118–126, 2013.

[27] Serena Yeung, Anitha Kannan, Yann Dauphin, and Li Fei-Fei. Tackling Over-pruning in Variational Autoencoders. jun 2017.

Supplementary Material

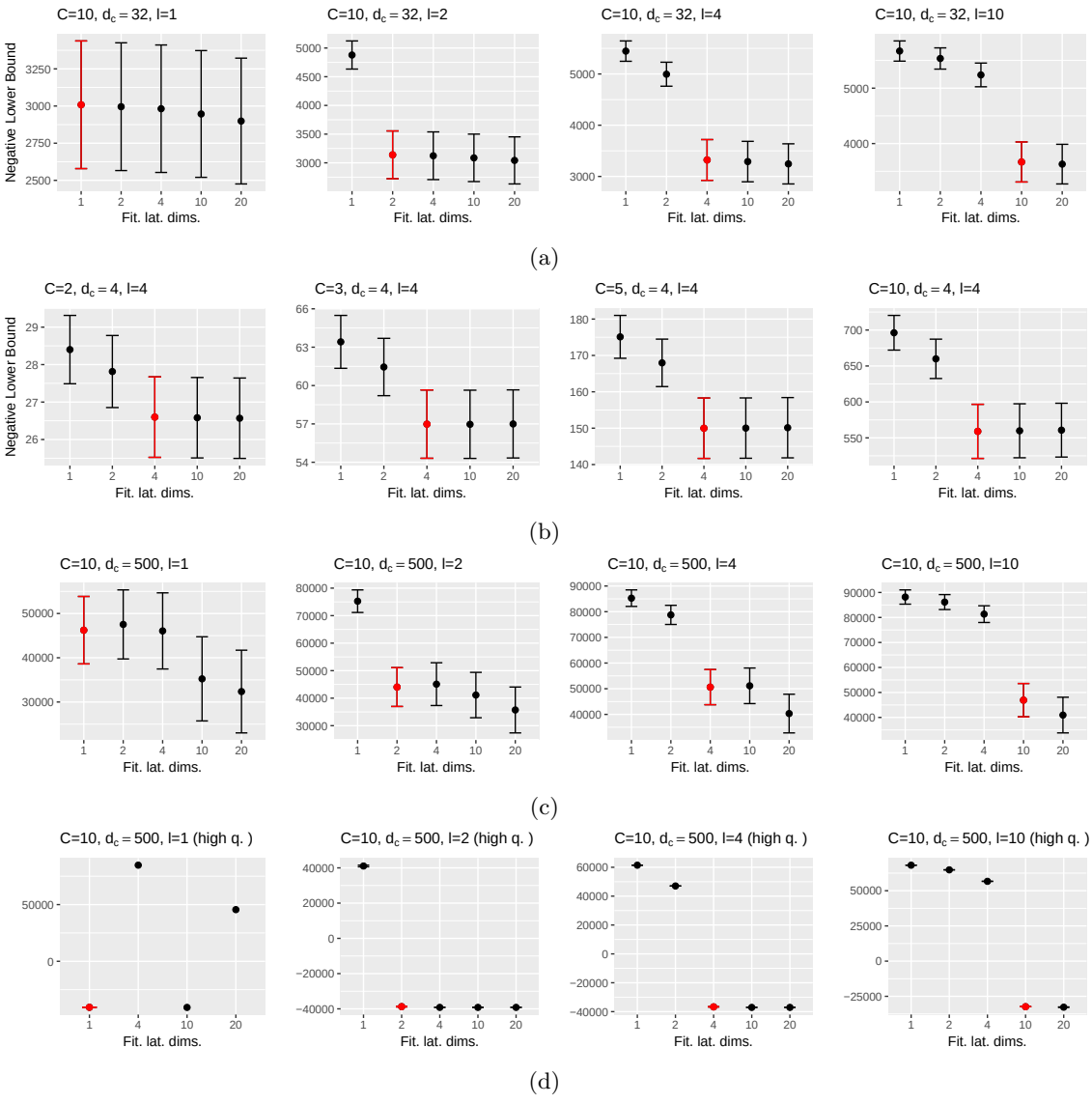
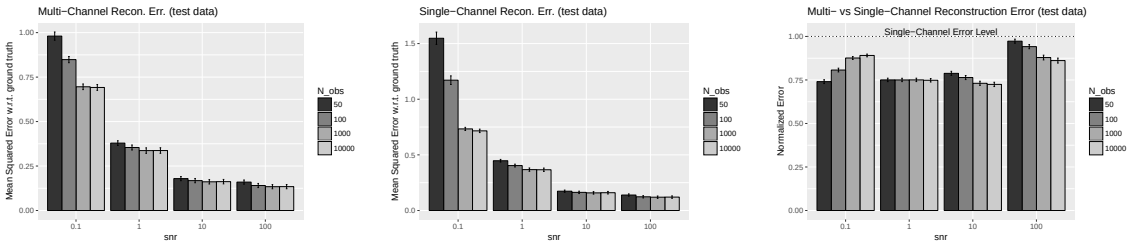


Figure S1: Negative lower bound (NLB) on the synthetic training set computed at convergence for all the scenarios. Each bar shows mean ± std.err. of $N = 80$ total experiments as a function of the number of fitted latent dimensions. Red bars represents experiments where the number of true and fitted latent dimensions coincide. (a) Experimental setup $C = 10$, $d_c = 32$: NLB stops decreasing when the number of fitted latent dimension coincide with the generated ones; notable gap between the under-fitted and over-fitted experiments (elbow effect). (b) Experimental setup $d_c = 4$, $l = 4$: increasing the number of channels C makes the elbow effect more pronounced. (c) Experimental setup $C = 10$, $d_c = 500$: with high dimensional data ($d_c = 500$) using the lower bound as a model selection criteria to assess the true number of latent dimensions may end up in overestimation. (d) Restricted ($N = 5$ total experiments) high quality experimental setup $C = 10$, $d_c = 500$, $S = 10000$, $snr = 100$: the risk to overestimate the true number of latent dimensions can be mitigated by increasing the snr and S of the observations in the dataset.



(a) (b) (c)

Figure S2: Reconstruction error on synthetic test data reconstructed with the multi-channel model. The reconstruction is better for high *snr* and high training data sample size. Scenarios where generated by varying one-at-a-time the dataset attributes listed in Tab. 1 for a total of 8 000 experiments. (a) Mean squared error from the ground truth test data using the Multi-Channel reconstruction: $\hat{\mathbf{x}}_i = \mathbb{E}_c \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)} \left[p\left(\mathbf{x}_i | \mathbf{z}, \boldsymbol{\theta}_i\right) \right] \right]$. (b) Mean squared error from the ground truth test data using the Single-Channel reconstruction: $\hat{\mathbf{x}}_i = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\phi}_i)} \left[p\left(\mathbf{x}_i | \mathbf{z}, \boldsymbol{\theta}_i\right) \right]$. (c) Ratio between Multi- *vs* Single-Channel reconstruction errors: we notice that the error made in ground truth data recovery with multi-channel information is systematically lower than the one obtained with a single-channel decoder.

Large Scale Cardiovascular Model Personalisation for Mechanistic Analysis of Heart & Brain Interactions

Jaume Banus[1], Marco Lorenzi[1], Oscar Camara[2] and Maxime Sermesant[1]

1. Université Côte d'Azur, INRIA Sophia Antipolis, EPIONE research group, France

2. PhySense, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain

Originally published in:
Functional Imaging and Modeling of the Heart (FIMH), 2019.

Abstract

Cerebrovascular diseases have been associated with a variety of heart diseases like heart failure or atrial fibrillation, however the mechanistic relationship between these pathologies is largely unknown. Until now, the study of the underlying heart-brain link has been challenging due to the lack of databases containing data from both organs. Current large data collection initiatives such as the UK Biobank provide us with joint cardiac and brain imaging information for thousands of individuals, and represent a unique opportunity to gain insights about the heart and brain pathophysiology from a systems medicine point of view. Research has focused on standard statistical studies finding correlations in a phenomenological way. We propose a mechanistic analysis of the heart and brain interactions through the personalisation of the parameters of a lumped cardiovascular model under constraints provided by brain-volumetric parameters extracted from imaging, i.e: ventricles or white matter hyperintensities volumes, and clinical information such as age or body surface area. We applied this framework in a cohort of more than 3000 subjects and in a pathological subgroup of 53 subjects diagnosed with atrial fibrillation. Our results show that the use of brain feature constraints helps in improving the parameter estimation in order to identify significant differences associated to specific clinical conditions.

1 Introduction

Cerebrovascular diseases are related to a variety of heart diseases such as heart failure [1] or atrial fibrillation (AF) [2], sharing several risk factors such as cholesterol, diabetes or high blood pressure. In parallel, it has been shown that stroke doubles the risk of dementia [3]. All these connections suggest a common underlying pathological process that links cardiac function with brain atrophy. Large scale analysis on databases combining cardiovascular and brain data from the same individuals are thus required to demonstrate and better understand the interaction between brain and heart. To this end, studies such as the UK Biobank aim at the acquisition of multi-modal databases containing both heart and brain imaging information [4]. Thanks to these databases ongoing studies have focused on the study of the relationships between cardiovascular risk factors

and image-derived features, such as subcortical volumes [5]. However, a number of fundamental descriptors of the cardiac function are not possible to obtain in-vivo, i.e: heart contractility or fibers stiffness. Personalised modelling approaches allow us to estimate these descriptors and gain insight of the cardiac function, allowing us to obtain more reliable results and relate them to brain damage information.

Personalizing a cardiovascular model for a given subject is an ill-posed problem that implies estimating the model parameters so that the simulation behaves as close as possible to the available clinical data. In this work we will focus on a 0D model of the whole cardiovascular system. Previous studies have used multi-scale models to describe the whole-body circulation and study the venous blood flow in the brain [6]. However, their modeling of the heart chambers does not take into account the contractile and elastic properties of the heart. While other models of the whole-body circulation are available [7, 8], to our knowledge, no explicit modelling study relating neurodegeneration and cardiovascular parameters has been done.

In this paper we aim to study the relationship between cardiovascular indicators and brain volumetric features extracted from the imaging data available in UK Biobank, through the personalisation of a cardiovascular lumped model using the approach presented in [9]. The use of this approach allows us to tackle the ill-posedness nature of the personalisation and identify plausible and coherent solutions across the population. To achieve that, we define a regularisation term that can be extended to take into account features not present in the lumped model, allowing to explore the effect of including brain features as additional constraint. We apply this framework to a large cohort composed by more than 3 000 subjects for which cardiac and brain information was jointly available in the UK Biobank. To illustrate how to exploit the framework to identify meaningful clinical relationships, we applied it in a subset of subjects diagnosed with AF, which is considered as an independent risk factor for stroke and dementia [2, 10]. We identified statistically significant associations between the personalised model parameters and brain volumetric features that match findings reported in previous clinical studies.

The paper is structured as follows: in section 2.1 we detail the data pre-processing and inclusion criterion for the whole-population analysis. Following, in section 2.2 we present the lumped model and how to take into account the subject's information to constrain the solution space in the personalisation. Next, in section 2.3 we assess the impact of our model in determining significant relationships between the estimated cardiac parameters and brain damage using the AF subset. Finally, in section 3 we present the obtained results.

2 Methods

2.1 Data pre-processing and inclusion criterion

Our analysis includes data from UK Biobank participants for which all brain image modalities and all cardiac-image derived indicators were available, for a total of 3783 subjects. In the available cardiac images it was possible to quantify the cardiac function using indicators such as stroke volume (SV), cardiac output (CO) or ejection fraction (EF). Multi-modal brain MRI images allowed the extraction of image-derived features such as brain tissue volumes and white matter hyperintensities (WMHs), one of the most common indicators used to assess neurological damage.

Using FLAIR MR images, WMHs were segmented by the lesion prediction algorithm (LPA),

available in the lesion segmentation toolbox (LST) [11] for SPM [1]. FLAIR MR images were pre-processed following the protocol described in [12], in which gradient distortion correction and defacing were performed. After discard subjects for which pre-processing (449) or segmentation of WMHs (250) failed, the final number of available subjects was 3 084. From the segmentations we extracted the total volume of WMHs and the number of lesions. All brain-related volumes were normalized by head size.

2.2 Cardiovascular lumped model

The cardiovascular personalisation of the subjects was performed by using the 0D model shown in Figure 1 which is a simplification of a 3D cardiac electromechanical model [13] derived in [14]. In the 0D version, which assumes spherical symmetry, the myocardial forces and motion can be described by the inner radius (R_0) of the ventricle. Deformation and stress tensors are also reduced to 0D forms, which allow us to characterise the heart contractile (σ_0) and elastic (C_1) properties of the heart.

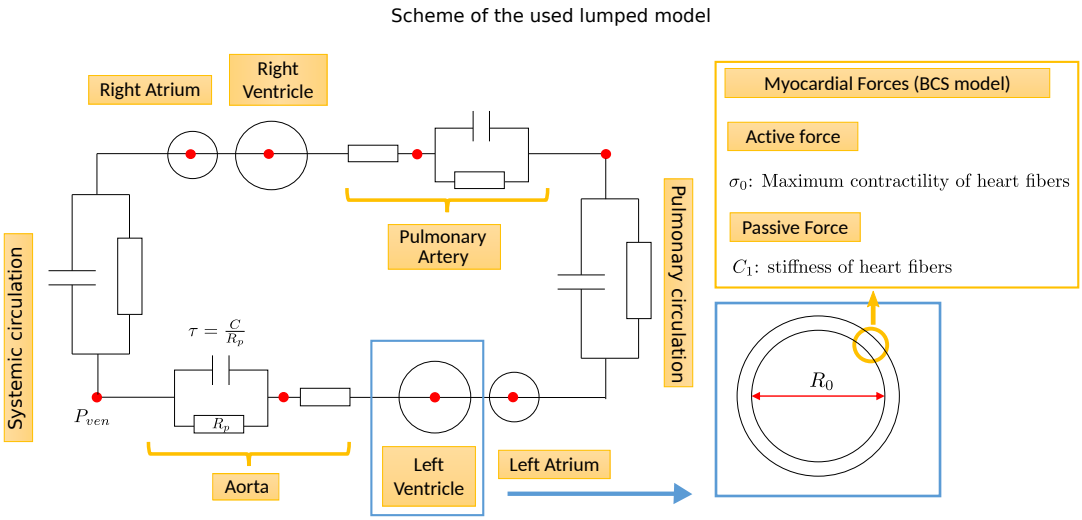


Figure 1: Simplified schematic representation of the lumped model showing the parameters used in the personalisation. The 0D representation of the myocardial forces has been omitted for the sake of clarification. τ characterizes the contractility of the aorta, R_p the peripheral resistance and P_{ven} the venous pressure.

The model M consists in a set of ordinary differential equations with P_M parameters, e.g. maximum contraction of the heart fibers or its stiffness. The state variables of the model are denoted by O_M, e.g. arterial or venous pressures, and they describe the state of the system. During the personalisation we are interested in a subset of n state variables, such that $O = (O_1, O_2, ..., O_n)$, and we vary a subset θ of the P_M model parameters. We consider $O(\theta)$ the set of state variables

[1] https://www.fil.ion.ucl.ac.uk/spm

generated by the model for a given set of θ. The goal is to find $\theta*$ such that $O(\theta*)$ best approximates the target features $\hat{O}$.

Due to the high dimensionality and non-convexity of this inverse problem, we solve it with the CMA-ES optimization algorithm based on evolution strategies [15]. CMA-ES minimizes a given error function by combining maximum likelihood principles with natural gradient descent on the ranks of the point scores (i.e: the score of each individual at every generation). The error function $S(\theta, \hat{O})$ is defined as the L_2 distance between $O(\theta)$ and $\hat{O}$. Since each target feature has different range of values we defined a tolerance interval, Tol, for each feature i to be able to compare the different outputs. This can be formalized as shown in Equation 1:

$$S(\theta, \hat{O}) = \sum_{i=1}^{n} \frac{(O_i(\theta) - \hat{O}_i)^2}{Tol_i} \tag{1}$$

Based on the available clinical data, we selected the following target features for the personalisation; stroke volume (SV), ejection fraction (EF), diastolic blood pressure (DBP), mean blood pressure (MBP) and end-diastolic volume (EDV). Considering the uncertainty of the measured data, the tolerance interval for each feature was set to 10 ml for the SV and the EDV, 200 Pa for the DBP and the MP, and 5% for the EF. Finally, the personalized parameters of the cardiovascular model were maximum contraction of the heart fibers σ_0, stiffness of the heart fibers C_1, peripheral resistance R_p, venous pressure P_{ven}, and the characteristic time τ of the aorta, which defines the time that takes for blood pressure to decrease from systolic to the systemic, or "asymptotic" value. We selected these parameters based on a sensitivity analysis in which we assessed the influence of each parameter over the selected target features.

Since the solution of equation 1 is non-unique, there is an observability difficulty in this personalisation problem. To tackle this issue, we used the iterative-update prior (IUP) approach presented in [9] to introduce constraints in the fitting process. In the IUP method a regularization term, $R(\theta, \mu, \Sigma)$, is used to reduce the variability in the estimation of the parameters. The regularization constrains the directions in which we explore the parameter-space by using the relationships among the model parameters. Formally, the regularization term is parameterized by an expected value μ and by a covariance matrix Σ encoding the relationships across parameters.

$$R(\theta, \mu, \Sigma) = (\theta - \mu)^T \Sigma^{-1} (\theta - \mu). \tag{2}$$

Therefore, the fitting score becomes:

$$S(\theta, \hat{O}, \mu, \Sigma) = S(\theta, \hat{O}) + \gamma R(\theta, \mu, \Sigma), \tag{3}$$

where γ defines the relative importance of the regularization term. This term is updated at each IUP iteration, using the obtained mean value of the fitted parameters and the estimated covariance in the previous iteration.

2.2.1 Accounting for brain information in the 0D model.

$R(\theta, \mu, \Sigma)$ can be extended to incorporate relationships with features not present in the cardiovascular model. In our setting, we included in the regularization term the extended feature space corresponding to the concatenation of the model parameters, θ, with the brain and clinical information, here denoted by ϕ. We used the total brain volume, the ventricles volume, the obtained

WMHs features, age, sex and body surface area (BSA). Therefore, the problem in equation 3 becomes:

$$S(\theta, \hat{O}, \mu, \Sigma, \phi) = S(\theta, \hat{O}) + \gamma R(\theta, \phi, \mu, \Sigma). \tag{4}$$

Equation 4 now accounts for a covariance term constraining the parameters according to the extended set of information. We have used 10 IUP iterations and assessed the results at different γ levels (0.1, 0.5, 2 and 10). The optimisation is performed over the logarithm of the parameter values.

2.3 Atrial fibrillation analysis.

Considering the dataset obtained after the pre-processing described in section 2.1 we had access to 53 subjects diagnosed with AF. Using bootstrapping we sampled 100 control groups of the same sample size of the AF group and without any significant difference in age, sex and BSA. The sampled controls came from the subset of subjects without any diagnosed cardiovascular disease ($n = 2022$). We applied the framework described in the previous section to each bootstrap subset composed by the AF group and sampled control group, to obtain the bootstrap distributions of the correlations between cardiac and external parameters. This approach allowed us to exploit the dataset variability for assessing the difference between cardiac and brain associations.

3 Results

3.0.1 Whole-population analysis.

As expected, we observed that as the value of γ is increased (i.e. more regularisation), the fitting error increases and at the same time the number of outliers is reduced and the estimated distributions have lower variability, as can be seen in Figure 2a. We can observe that strong regularisation even shrinks some parameters close to a constant value, implying that those parameters cannot be observed from the available data. Looking at the correlation of the model parameters with the external features we note the strong correlation between the left ventricle size, R_0, and the BSA, even for low γ values. Moreover, there is a positive correlation between peripheral resistance, the WMHs volume, and brain ventricles volume, which are at the same time negatively correlated to the aorta characteristic time, τ. The number of WMHs lesions and age followed the same correlations pattern, but due to space issues they have been omitted. On the other hand, brain volume is positively correlated with τ and peripheral resistance. An increase in peripheral resistance can be associated to higher DBP, while a decrease in contractility, τ, can be interpreted as an increase in arterial stiffness leading to high SBP. Both, DBP and SBP, have been previously associated to WMHs [16]. In figure 2b we note that the significant correlations present when no regularization is applied become stronger as regularization increases, while the non-significant correlations stay close to zero. This behavior is expected since regularization is constraining the space of feasible solutions. Therefore, as we increase γ we further limit the feasible parameter-space towards the set of solutions that satisfy the existing relationships between the parameters. In figure 2c we observe the obtained correlations when the external features are not taken into account. In this case, the solutions are constrained into a different parameter-space in which the relationships between the model parameters and the external features are lost. Limiting the interpretability when assessing the parameters estimations with respect external factors not present in the mechanistic model.

3.0.2 Atrial fibrillation.

In Figure 3 we observe a statistical description of the empirical distribution of correlations obtained from the bootstrap analysis done in the AF subset. The results are obtained with trade-off $\gamma = 0.5$, which in the whole-population analysis provided a good a balance between data-fit and regularization. We assessed the difference between the controls and AF groups correlations distributions using the Wilcoxon rank-sum test with a significance level of $\alpha = 0.05$. Moreover, to consider the obtained results as significant we assessed if in the AF group the obtained correlations were statistically greater or smaller than 0 with a 5% significance level.

In the brain volume we observe the same correlations found in the whole-population analysis, but it can be seen that for the AF group these correlations are stronger, which suggests that in the AF subjects brain is more susceptible to cardiovascular factors. For the BSA we found the positive correlation with the left ventricle size observed in section 3.0.1 and a negative correlation with the maximum heart contractility σ_0. In the number of WMHs lesions we can observe a positive correlation with the left ventricle size. The associations of BSA with σ_0 and WMHs with R_0 could be related to cardiac dilation due to an increased impairment of the functioning heart in AF. Interestingly, in the AF control subjects the correlation between WMHs and left ventricle size is negative. These findings suggest an association between AF and WMHs. Moreover, they agree with previous studies reporting an association between left ventricle remodelling and AF [17]. No significant associations were found for brain ventricles volume, WMHs volume and age.

4 Conclusions

We have modeled 3 084 subjects with a 0D cardiovascular model and we constrained the available parameter-space during personalisation by incorporating external features in the regularization term, allowing us to study their influence in the estimated model parameters. The use of this approach gives access to a generative model that allows to analyze the relationships between external features and non-observable parameters such as the characteristic time of aorta, τ, which we found to be related with brain-volumetric features. Using the same framework we assessed a clinical subgroup in which we have found meaningful clinical relationships, linking AF with WMHs and heart remodelling. Our model does not currently simulate the cerebral blood flow, while previous studies [6] suggest that WMHs are due to more localize vascular impairments. This highlights the need to obtain a local flow characterization to estimate more relevant parameters. Moreover, the presented approach can be seen as a parameter selection approach. It allows to identify which parameters cannot be estimated from the available data and find a parameter subspace of solutions in which the non-observable parameters get close to constant values. The identification of the non-observable parameters coupled with human modelling expertise can help in the selection of a reduced subset of observable cardiovascular parameters for personalisation. Future work will go towards the local blood flow characterization in the brain, as well as towards the assessment of its spatial patterns, and the modelling of more brain atrophy indicators.

4.0.1 Acknowledgements.

This work was supported by the Inria Sophia Antipolis - Mditerrane, "NEF" computation cluster. This research has been conducted using the UK Biobank Resource undder Application Number 20576 (PI Nicholas Ayache). Additional information can be found at: https://www.ukbiobank.ac.uk

References

[1] A. Ois et al. Heart failure in acute ischemic stroke. *Journal of Neurology*, 255(3):385–389, 2008.

[2] E. J. Benjamin et al. Heart disease and stroke statistics2018 update: A report from the american heart association. *Circulation*, 137(12), 2018.

[3] M. R. Azarpazhooh et al. Concomitant vascular and neurodegenerative pathologies double the risk of dementia. *Alzheimers and Dementia*, 14(2):148–156, 2018.

[4] C. Sudlow et al. Ukbiobank an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12:1–10, 2015.

[5] Simon R Cox et al. Associations between vascular risk factors and brain MRI indices in UK Biobank. *bioRxiv*, 2019.

[6] L. O. Mller and E. F. Toro. Enhanced global mathematical model for studying cerebral venous blood flow. *Journal of Biomechanics*, 47(13):3361–3372, 2014.

[7] S. Safaei et al. Bond graph model of cerebral circulation: Toward clinically feasible systemic blood flow simulations. *Frontiers in Physiology*, 9:1–15, 2018.

[8] P. Blanco et al. An anatomically detailed arterial network model for one-dimensional computational hemodynamics. *IEEE Transactions on Biomedical Engineering*, 62(2):736–753, 2015.

[9] R. Mollro, X. Pennec, H. Delingette, N. Ayache, and M. Sermesant. Population-based priors in cardiac model personalisation for consistent parameter estimation in heterogeneous databases. *International Journal for Numerical Methods in Biomedical Engineering.*, 2018.

[10] Alvaro Alonso and Antonio P Arenas de Larriva. Atrial Fibrillation, Cognitive Decline and Dementia. *European Cardiology Review*, 11(1):49, 2016.

[11] Paul Schmidt. Bayesian inference for structured additive regression models forlarge-scale problems with applications to medical imaging. *PhD thesis, Ludwig-Maximilians-Universitt Mnchen*, 2017.

[12] F. Alfaro-Almagro et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *NeuroImage*, 166:400–424, 2018.

[13] D. Chapelle, P. Le Tallec, P. Moireau, and M. Sorine. Energy-preserving muscle tissue model: formulation and compatible discretizations. *International Journal for Multiscale Computational Engineering*, 10(2):189–211, 2012.

[14] M. Caruel, R. Chabiniok, P. Moireau, Y. Lecarpentier, and D. Chapelle. Dimensional reductions of a cardiac model for effective validation and calibration. *Biomechanics and Modeling in Mechanobiology*, 13:897–914, 2014.

[15] N. Hansen. The cma evolution strategy: A comparing review. *Towards a New Evolutionary Computation*, 102:75–102, 2016.

[16] R. Modir, H. Gardener, and C. B. Wright. Stroke blood pressure and white matter hyperintensity volume a review of the relationship and implications for stroke prediction and prevention. *US Neurology*, 8(1):33–36, 2012.

[17] Y. Seko et al. Association between atrial fibrillation, atrial enlargement, and left ventricular geometric remodeling. *Scientific Reports*, 8(1):1–8, 2018.

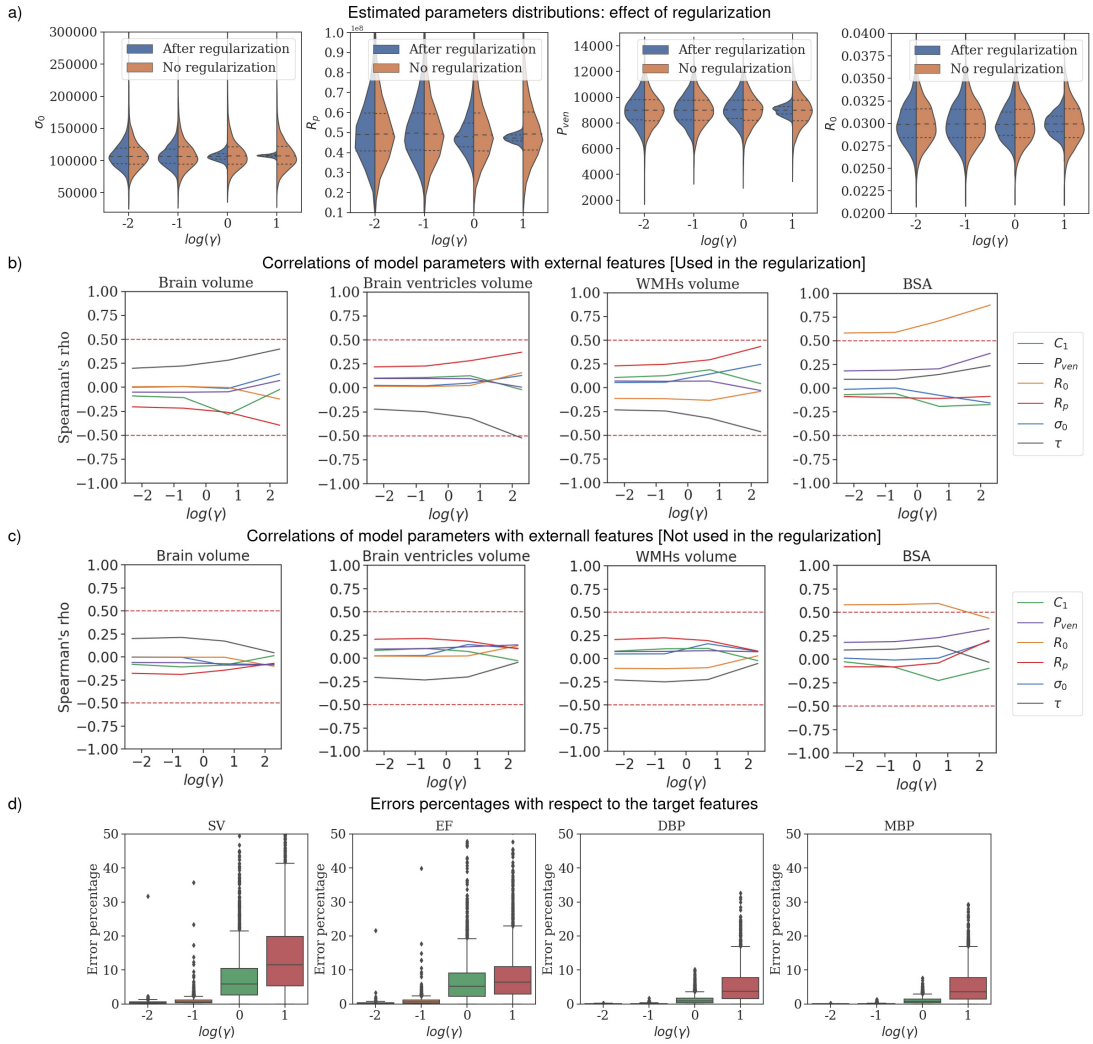


Figure 2: a) Estimated density distributions of the fitted parameters at different regularization levels. Initial and final distributions after 10 iterations in brown and blue respectively. The variability among the initial distributions is due to the variability in the sampling that CMA-ES performs during the optimization.Error b) Evolution of Spearman's rank correlation coefficient between the model parameters and the external parameters as the regularization level increases when external features are considered in the regularization and c) when external features are not considered. Model parameters being: maximum heart fibers contractility σ_0, heart fibers stiffness, C_1, left ventricle size R_0, peripheral resistance R_p, aorta characteristic time τ, and venous pressure P_{ven}. d) Error percentages with respect to the target features. Stroke volume (SV), ejection fraction (EF), diastolic blood pressure (DBP) and mean blood pressure (MBP). While end-diastolic volume (EDV) is not shown due to space issues, its error pattern was similar to the one observed in the SV

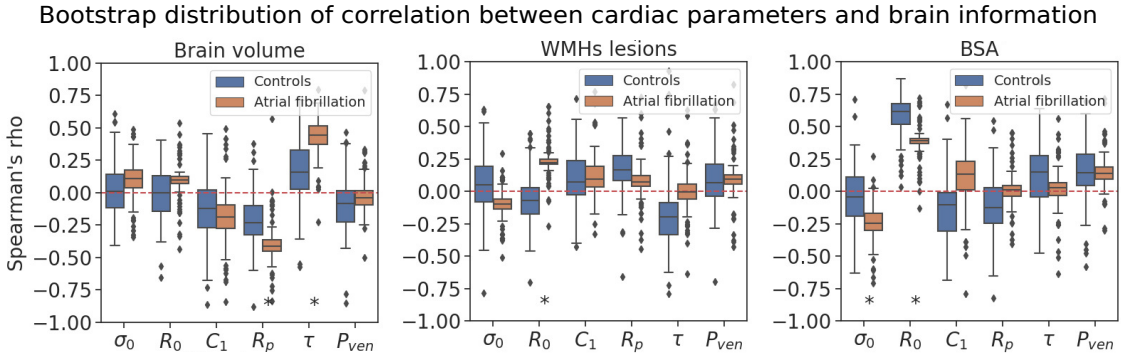


Figure 3: Comparison of the Spearman's rank correlation coefficient bootstrap distributions obtained at $\gamma = 0.5$ between the personalised model parameters and the external features. Blue boxplots correspond to control groups and brown to AF subjects. $*$ denotes that the correlations are significantly different according to the Wilcoxon rank-sum test, and that in the AF group correlations are significantly greater or smaller than 0 (5% significance level).

Susceptibility of brain atrophy to *TRIB3* in Alzheimer's disease, evidence from functional prioritization in imaging genetics

Marco Lorenzi[a,b,1,2], Andre Altmann[a,1], Boris Gutman[c], Selina Wray[d], Charles Arber[d], Derrek P. Hibar[c], Neda Jahanshad[c], Jonathan M. Schott[e], Daniel C. Alexander[f], Paul M. Thompson[c], and Sebastien Ourselin[a], for the Alzheimer's Disease Neuroimaging Initiative[3]

[a]Translational Imaging Group, Centre for Medical Image Computing, University College London, London WC1E 7JE, United Kingdom; [b]Epione Research Project, Université Côte d'Azur, BP 93 06 902 Inria Sophia Antipolis, France; [c]Imaging Genetics Center, University of Southern California, Marina Del Rey, CA 90292; [d]Department of Molecular Neuroscience, University College London Institute of Neurology, London WC1N 1PJ, United Kingdom; [e]Department of Neurodegeneration, Dementia Research Centre, Institute of Neurology, London WC1N 3BG, United Kingdom; and [f]Centre of Medical Image Computing, University College London, London WC1E 7JE, United Kingdom

Edited by Marcus E. Raichle, Washington University in St. Louis, St. Louis, MO, and approved January 31, 2018 (received for review April 19, 2017)

The joint modeling of brain imaging information and genetic data is a promising research avenue to highlight the functional role of genes in determining the pathophysiological mechanisms of Alzheimer's disease (AD). However, since genome-wide association (GWA) studies are essentially limited to the exploration of statistical correlations between genetic variants and phenotype, the validation and interpretation of the findings are usually nontrivial and prone to false positives. To address this issue, in this work, we investigate the functional genetic mechanisms underlying brain atrophy in AD by studying the involvement of candidate variants in known genetic regulatory functions. This approach, here termed functional prioritization, aims at testing the sets of gene variants identified by high-dimensional multivariate statistical modeling with respect to known biological processes to introduce a biology-driven validation scheme. When applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, the functional prioritization allowed for identifying a link between tribbles pseudokinase 3 (*TRIB3*) and the stereotypical pattern of gray matter loss in AD, which was confirmed in an independent validation sample, and that provides evidence about the relation between this gene and known mechanisms of neurodegeneration.

imaging–genetics | Alzheimer's disease | brain atrophy | *TRIB3* | neuroimaging

Alzheimer's disease (AD) is a devastating neurodegenerative disorder, and its etiology still remains largely concealed. In anticipation of increasing prevalence of AD and other dementias, there is an urgent need for improving the understanding of the disease processes that underlie neurodegeneration. While the knowledge about the genetic and environmental risks underpinning AD is steadily advancing, how these factors interact to lead to the complex pathophysiology that results in dementia is less understood.

Advances in imaging technologies have led to noninvasive or minimally invasive imaging biomarkers that capture various aspects of the disease process, including amyloid deposition (1), tau pathology (2), functional decline (3), and neuronal loss (4). Combining such imaging information with genetic measurements—so-called imaging–genetics—provides the means for investigating the effect of genetic variation on underlying biological mechanisms (5).

Genome-wide association studies (GWAS) query millions of SNPs individually for their association with either case–control status (6) or disease-specific quantitative phenotypes [e.g., in the case of AD, regional brain volumes (7) or brain amyloid burden (8)]. Mass univariate analysis of genetic data is still the predominant method in virtue of its ease of use and well-established theoretical framework, albeit that it suffers from significant limitations, including the requirement for multiple testing, redundancies introduced by linkage disequilibrium (LD), and the lack of analysis of epistatic effects (e.g., SNP–SNP interactions), which have to be explicitly modeled and searched for exhaustively (9). Moreover, more than one quantitative phenotype can be derived from the available imaging data (e.g., dozens or hundreds of regional brain volumes or hundreds of thousands of voxel-level metrics) (10). This potentially large number of genotype–phenotypes features of interest generally complicates the problem of reliably detecting statistical associations and thus, hampers the identification of disease-relevant genetic markers by purely statistical means.

Limitations of classical mass univariate statistical methods have, in recent years, been overcome by using multivariate approaches to data analysis in the context of neuroscience studies (11) and GWAS (12). Likewise, in imaging–genetics, meaningful genotype–phenotype interactions (13) are captured by simultaneously

Significance

In this study, we use an experimental imaging–genetics approach for investigating the genetic underpinnings of brain atrophy in Alzheimer's disease. We successfully combined state-of-the-art imaging–genetics methods and experimental gene expression data to uncover biology in brain atrophy. The experimental paradigm highlighted a significant role of tribbles pseudokinase 3 (*TRIB3*) in modulating the typical pattern of Alzheimer's brain pathology. This result corroborates through rigorous data-driven statistical methods evidence emerging from previous studies about the role of *TRIB3* in modulating known mechanisms of neurodegeneration, such as neuronal death, cellular homeostasis, and interaction with established genes causing autosomal dominant Alzheimer's disease: *APP* and *PSEN1*. The developed integrated statistical–experimental methodology could serve as a roadmap for investigations in other disorders.

Author contributions: M.L., A.A., S.W., D.P.H., N.J., J.M.S., D.C.A., P.M.T., and S.O. designed research; M.L., A.A., and C.A. performed research; M.L., A.A., and B.G. analyzed data; and M.L., A.A., S.W., C.A., and J.M.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All the study data are available through the Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.usc.edu/). The experimental material, including PLS components and probability measures, has been deposited in Figshare (https://doi.org/10.6084/m9.figshare.5914375.v1).

[1]M.L. and A.A. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: marco.lorenzi@inria.fr.

[3]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1706100115/-/DCSupplemental.

Published online March 6, 2018.

modeling sets of genetic variants that are jointly associated with a given imaging phenotype (14–17). Multivariate GWAS have the potential to shed light on the complex genotype–phenotype relationship and may thus highlight novel links between brain physiology and molecular and biological functions. However, although these methods have proven their ability to identify meaningful SNP combinations associated with brain imaging features, the interpretation and validation of the statistical findings remain very challenging tasks. These problems relate directly to the understanding of the functional role of sets of genetic variants and to the difficulty of replicating the statistical results in unseen cohorts.

We approach this technical bottleneck by leveraging multivariate approaches to explore high-dimensional datasets and to generate hypotheses, which are subsequently tested in downstream experiments. High-quality databases of matched genotype and gene expression measurements, such as the Genotype Tissue Expression project (GTEx; gtexportal.org/home/) (18) and BRAINEAC (www.braineac.org/) (19), facilitate the quantification of effects of SNPs on gene expression in numerous tissues, including various brain tissues. Typically, these databases are used to detail the effect of a genetic variant at the very end of an analysis pipeline and to garner evidence for molecular mechanisms of the genetic locus. However, functional information in "convenience" databases can also be used at an earlier stage in the analysis to prioritize a few candidate hypotheses with a clear functional mechanism [e.g., expression quantitative trait loci (eQTL)] for the validation phase and thus, limit the multiple testing burden.

In this work, we apply this investigative approach to study the genetic functional mechanisms underlying brain atrophy in AD. The framework is composed of two steps.

i) Statistical discovery. Candidate genetic variants are initially identified through data-driven multivariate statistical analysis of the matched imaging and genetics data (Fig. 1). This is achieved by modeling the joint covariation between 1.1 million SNPs and the cortical and subcortical atrophy represented by 327,684 cortical and 27,120 subcortical thickness values of 639 individuals (either healthy older controls or patients with AD) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort.

ii) Functional prioritization. The candidate genetic variants are subsequently screened for functional relevance by querying high-dimensional gene expression databases, such as GTEx.

The resulting small set of genetic loci, which are shown to modify gene expression, is then validated in an independent sample of 553 individuals from the ADNI diagnosed with mild cognitive impairment (MCI), a proportion of whom progressed to AD.

Compared with previous approaches, our work (*i*) analyzes the whole genome and whole brain in a hypothesis-free fashion (i.e., without preselecting SNPs or brain regions) and (*ii*) uses a functional prioritization step to select genetic loci for validation in an independent cohort.

Starting from the initial ∼1.1 million SNPs, the multivariate statistical analysis allowed for the identification of a relatively small number of genetic loci that are statistically associated with the typical pattern of AD brain pathology. The subsequent functional prioritization step ultimately identified a significant role of tribbles pseudokinase 3 (*TRIB3*), a gene showing important connections to known mechanisms of neurodegenerative diseases. Indeed, although a role for *TRIB3* in dementia has not been extensively explored, there are several aspects of *TRIB3* function that have relevance to mechanisms related to neuronal death, cellular homeostasis, and interaction with established AD genes, such as *APP* and *PSEN1*.

This study ultimately offers an illustration of the potential of effectively combining multivariate statistical modeling in imaging–genetics with recent instruments available from computational biology to lead to insights on the pathophysiology of neurodegeneration.

Results

Model Training and Estimated Components. Figs. 2 and 3 show the relevant areas of the identified joint genetic and phenotype

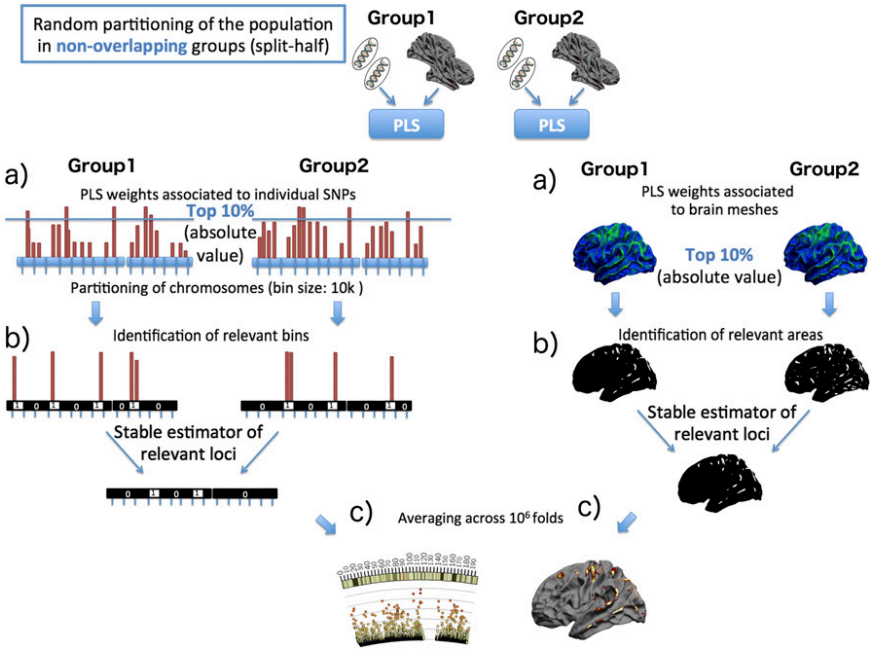


Fig. 1. Cross-validation scheme for the assessment of the genetic loci of maximal genotype–phenotype correlation identified by the PLS model. The whole procedure is repeated 1 million times, and the resulting array is further analyzed. (*A*) PLS is applied in a split-half setting. For each of the two nonoverlapping randomly sampled groups, the PLS components of joint phenotype and genotype variation are independently estimated. (*B*, *Left*) Each chromosome is partitioned in bins of 10-kb size, which are labeled one if they contain an SNP associated with the largest PLS weights (top 10% of absolute values) or zero otherwise. To obtain stable estimates of the loci of maximal weights, the resulting binary arrays independently estimated in the two groups are merged (binwise and operation). The same procedure is applied on the mesh-based PLS weights associated with the phenotype component. (*C*) Steps *A* and *B* are repeated across 1 million folds, and the results are subsequently averaged to obtain the confidence maps associated with genetic and phenotype components (Figs. 2 and 3).

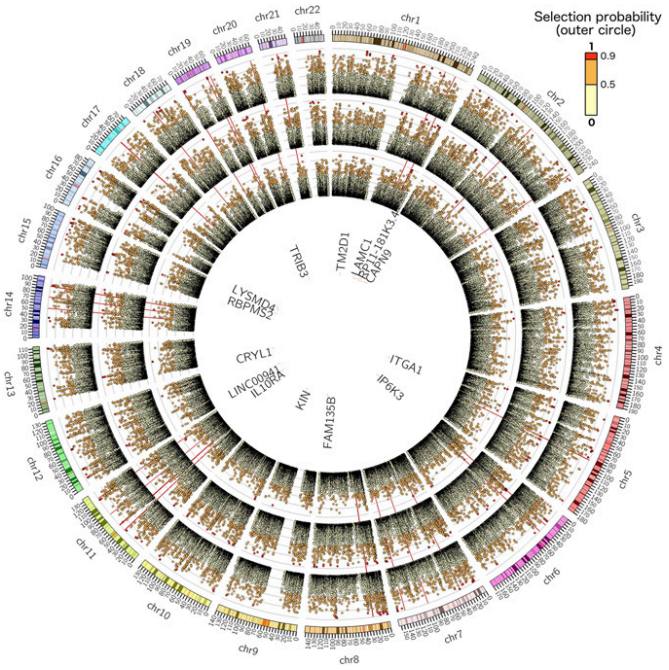


Fig. 2. PLS genotype component: the circular plots show the probability of a given genetic locus being associated with the phenotype components shown in Fig. 3. Outer to inner circles represents components 1–3, respectively. The plots show the probability of a given genetic bin of size 10 kb being relevant in the PLS model (i.e., to contain an SNP that is ranked in the top 10% of the absolute weights of the genotype component). The genes with eQTL close to the important loci ($P > 0.95$) are listed in the innermost circle depending on their genomic position. The red radial lines are located in correspondence of known AD genes: *ABCA7*, *APOE*, *APP*, *BIN1*, *CASS4*, *CD2AP*, *CD33*, *CELF1*, *CLU*, *CR1*, *DSG2*, *EPHA1*, *FERMT2*, *HLA-DRB5*, *INPP5D*, *MAPT*, *MEF2C*, *MS4*, *NME8*, *PICALM*, *PSEN1*, *PSEN2*, *PTK2B*, *SLC24A4*, *SORL1*, and *ZCWPW1*. High-resolution circular plots for each component are provided in Figs. S2–S4.

variation, respectively, for the first three partial least squares (PLS) components through stability selection. The components were very robust (100% reproducible) during the stability selection procedure (*SI Materials and Methods*). The fourth and fifth components did not present any relevant locations (i.e., all bins have $P <$ 0.95) after stability selection for both the genetic modality and the imaging modality.

Genetic Components. The circular Manhattan plot [Circos v0.96 (20)] of Fig. 2 shows the selection frequency for the PLS genotype components and describes the importance of the genetic loci associated with cortical thickness variation for components 1–3. The plot shows the probability of a given genetic bin of size 10 kb being relevant in the PLS model (i.e., to contain an SNP that is ranked in the top 10% of the absolute weights of the genotype component). Spatially contiguous loci generally show similar importance values, which are caused by LD of these regions.

In the genetic components 1–3, a total of 118 bins exceeded the selection frequency threshold (61, 50, and 7 for components 1–3, respectively). From these bins, 402 (196, 181, and 25) influential SNPs were extracted and annotated with 98 genes through the Ensembl Variant Effect Predictor (VEP) for GRCh37 (date accessed: October 17, 2016) (21). The extended Apolipoprotein E (*APOE*) locus comprising *APOE* and *TOMM40* was selected as the highest scoring region in component 1. A total of 3,956 candidate SNP–gene pairs were considered for the GTEx-based eQTL analysis in six tissues. However, a few genes did not show sufficient expression levels in some tissues, and these combinations were excluded from the analysis, resulting in 1,598 unique SNP–gene tissue tests, of which 104 were significant at the Bonferroni-corrected P value threshold ($P = 3.1e{-}5$) (Dataset S1) linking to 14 genes (Dataset S2 and Fig. S5): *CAPN9*, *CRYL1*, *FAM135B*, IL-10 receptor subunit alpha (*IL10RA*), *IP6K3*, *ITGA1*, *KIN*, *LAMC1*, *LINC00941*, *LYSMD4*, *RBPMS2*, *RP11-181K3.4*, TM2 domain-containing 1 (*TM2D1*), and *TRIB3*. These genes are listed in the innermost circle of Fig. 2 depending on their genomic position.

The independent validation of those 14 genes in the MCI cohort confirmed *TRIB3* ($P = 0.0034$) (Table 1). Three additional genes were close to nominal significance: *TM2D1* ($P = 0.053$), *LAMC1* ($P = 0.062$), and *RP11-181K3.4* ($P = 0.053$) (Table 1). Of note, the top eQTL SNP for *TRIB3* rs4813620 received a $P = 0.06175$ in stage 1 of a large AD GWAS (6). However, rs62191440, an SNP in strong LD with rs4813620 ($D' = 0.8469$; $r^2 = 0.6559$) in the European population (22), received a P value of 0.00601 (Fig. S6) and also constitutes an eQTL for *TRIB3* in various tissues in GTEx, including brain tissues cortex and caudate ganglia (Fig. S7). Interestingly, when estimating the PLS components on the subcohort of 279 training individuals positive to amyloid in the cerebrospinal fluid (CSF) (Table 2), we achieved compatible validation results in the independent MCI group. Within this setting, *TRIB3* still leads to marginally significant differences ($P = 0.0134$) between progressing and stable MCI (Table S1).

Morphometric Components. Fig. 3 shows the PLS phenotype components 1–3 along with the associated selection frequency describing the loci of brain atrophy associated with genetic variation. The first component is mainly associated with the thinning of the cortical mantle and is localized in temporal and posterior cingulate cortices. The relevant areas at the subcortical level are primarily associated with amygdalae and thalami. The second component is mostly associated with the thinning of the subcortical areas (hippocampi and amygdalae) and with the cortical thinning of the temporal areas at the cortical level. The third component is similar to component 2 and describes a subcortical thickness pattern prevalent in hippocampi, amygdalae, and thalami. At the cortical level, the component is associated with the thinning of frontal cortice and with isolated spots located in the parahippocampal gyrus.

Discussion

In this work, we modeled high-dimensional genome-wide SNP data and brain-wide cortical thickness data via joint multivariate statistical modeling and functional prioritization of genes through bioinformatics annotation and a large eQTL database.

Our study ultimately identified a link between *TRIB3* and the stereotypical pattern of gray matter loss in AD (cortical thinning in temporal and posterior cingulate regions and subcortical atrophy). *TRIB3* is a pseudokinase that acts as a regulator of several signaling pathways. For example, it can interact directly with Akt and inhibit the prosurvival Akt pathway (23). *TRIB3* expression is induced during neuronal cell death (24), and recently increased levels of the TRIB3 protein were found in dopaminergic neurons of the substantia nigra pars compacta in patients with Parkinson's disease (25). *TRIB3* expression is stress induced and increases in response to nerve growth factor deprivation, endoplasmatic reticulum stress, and amino acid deprivation (24). Although a role for *TRIB3* in dementia has not been extensively explored, there are several aspects of *TRIB3* function that have relevance to known mechanisms of neurodegenerative disease. TRIB3 can interact directly with P62 to modulate autophagic flux (26), an important process in maintaining cellular homeostasis that is known to be disrupted in neurodegeneration (27). Knockdown of *TRIB3* modulates PSEN1 stability (26), and a yeast two-hybrid screen identified progranulin as a direct interaction partner of TRIB3 (28). Intriguingly, it has recently been shown that TRIB3 induces both apoptosis and autophagy in Aβ-induced neuronal death, and silencing of *TRIB3* was strongly neuroprotective (29). These links warrant additional investigation for a functional role of *TRIB3* in neuronal death in dementia.

These earlier findings align with our eQTL analysis, where carriers of the minor allele show increased *TRIB3* expression (Fig. S5), which potentially lowers the threshold to *TRIB3*-mediated neuronal cell death. *TRIB3* expression was modulated by the identified SNP in various other tissues, including the caudate (Fig. S7), a region affected in Parkinson's disease and Huntington's

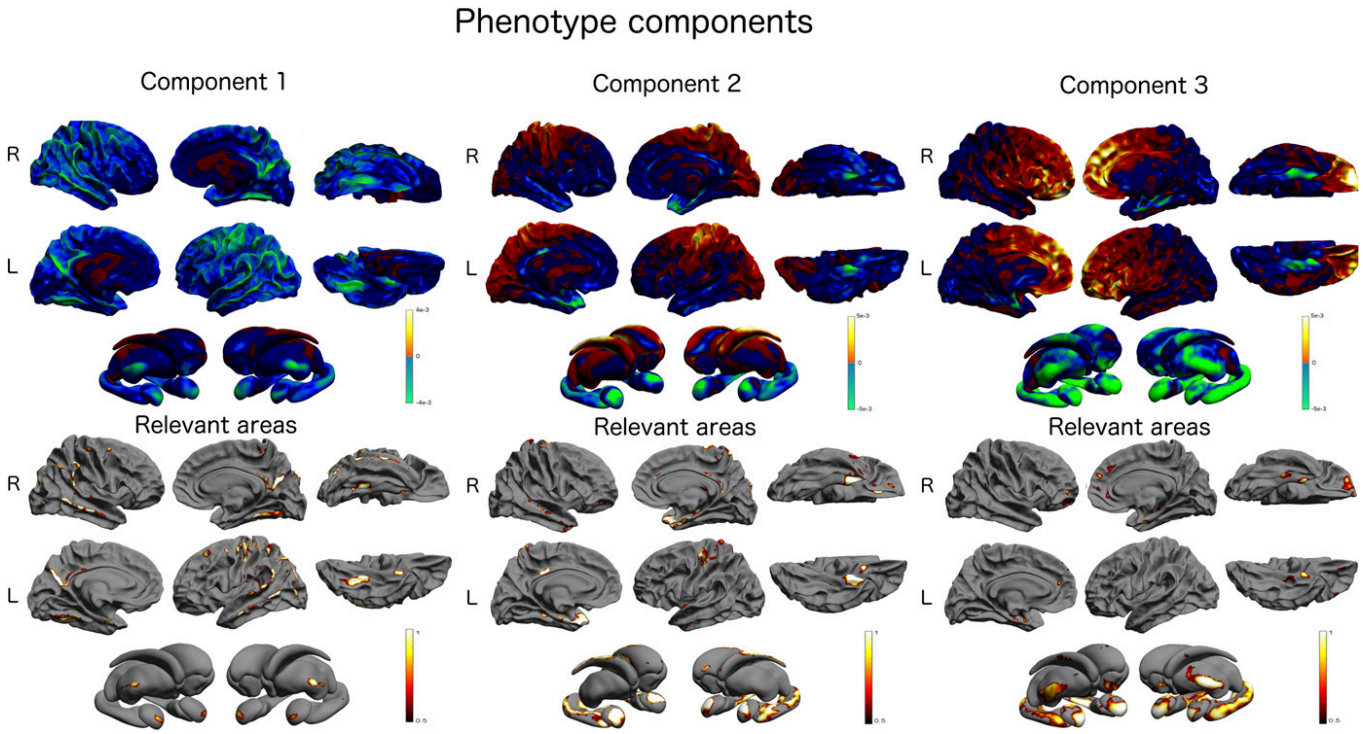


Fig. 3. PLS phenotype component: *Upper* shows the topographical distribution of the PLS weights associated with the cortical and subcortical brain areas. The absolute value of the weights is proportional to the importance of the underlying brain areas. The relevance of the brain areas is quantified in *Lower*. The colors (red to white) indicate the probability of a brain area being associated with the genotype component shown in Fig. 2 and quantify the probability of each cortical mesh point being relevant in the PLS model (i.e., to be ranked among the top 10% of the absolute weights of the phenotype component). L, left; R, right.

disease. A recent study of *Trib3* expression in mice concluded that "*Trib3* has a pathophysiological role in diabetes" (30); diabetes itself is a known risk factor for dementia (31), perhaps through shared metabolic processes with AD (32). Interestingly, one of three SNPs (rs1555318) selected in the PLS model and attributed to *TRIB3* showed a strong association with type 2 diabetes in stage 1 of a large GWAS (P = 4.4e−4) (Fig. S8) (33). Other GWAS showed links between *TRIB3* and information processing speed (P = 1.7e−7) (34) and AD (P = 0.006) (6). An earlier genetic study on AD in Swedish men found an association in *TRIB3* as well (P = 0.044) (35), which was replicated in a Canadian cohort (P < 0.001) (36). Lastly, TRIB3 was reported to physically interact with APP (37), and it shares numerous functional annotations for biological processes regarding lipid metabolism with APOE.

The functional prioritization component of the analysis successfully reduced the set of candidate genetic variants for the independent validation; however, this prioritization has a shortcoming: it hypothesizes that identified SNPs alter the expression of a nearby gene. Although this scheme led to the identification of *TRIB3* in the cortical thickness phenotype, it did miss a long-established AD risk gene: *APOE*. SNPs belonging to *APOE* (rs429358 and rs7412) were selected as the highest-scoring SNPs in component 1. However, none of them were detected as an eQTL, and thus, *APOE* was excluded from the downstream analysis. Other types of functional prioritizations based on exonic function prediction may have retained *APOE* and other genes in the pipeline. However, SNP data typically feature only a few nonsynonymous exonic variants, and their high frequency (minor allele frequency > 5%) renders them unlikely to receive significant "damaging" scores in these predictions. Thus, for this scenario, the use of these function predictions would be limited.

The list of genes that we identified contains other interesting candidates. For instance, *IL10RA* is a receptor for IL-10, a cytokine that controls inflammatory response (38). Carriers of the minor allele show increased *IL10RA* expression (Fig. S5), and *Il10ra* expression is increased in affected brain regions with increasing age and presence of AD pathology in transgenic mouse models of AD [MOUSEAC (39)] (Fig. S9). Moreover, a link between down-regulation of *IL10RA* and *TRIB3* in *TRIB3*-silenced HepG2 cells was reported in ref. 26 along with increased abundance of Presenilin 1, ApoE3, and Clusterin. Finally, blocking IL-10 response was recently suggested as a therapeutic mechanism in AD (40). A gene that showed a statistical trend in the validation sample was *TM2D1*, which is a beta-amyloid binding protein and may be involved in beta-amyloid–induced apoptosis (41). Furthermore, Myocyte Enhancer Factor 2A (*MEF2A*), like *APOE*, was filtered out by the functional prioritization. However, *MEF2A* is a paralog of *MEF2C*, which is an established AD gene (6). Noteworthy, bins covering *MEF2C* only

Table 1. Statistical comparison of the genes scores in training and testing groups (Kruskal–Wallis nonparametric test)

Gene	*P* value training (AD vs. healthy controls)	*P* value testing (MCI converter vs. MCI stable)
TM2D1	0.0050	0.0528
IL10RA	0.1069	0.6198
TRIB3	0.0032	0.0034
ZBTB7A	0.0360	0.9135
LYSMD4	0.0000	0.2057
CRYL1	0.6213	0.1176
FAM135B	0.0000	0.5588
IP6K3	0.0000	0.4646
ITGA1	0.0993	0.7310
KIN	0.0014	0.2061
LAMC1	0.0019	0.0618
LINC00941	0.0000	0.6896
RBPMS2	0.0000	0.2149
RP11-181K3.4	0.0017	0.0527

The score for TRIB3 leads to significant differences in the MCI testing group after Bonferroni correction for multiple comparisons.

Table 2. Sociodemographic, clinical, and genetic characteristics of the study cohort [mean (SD)]

	Discovery		Validation	
Diagnosis at imaging	Healthy	AD	MCI stable	Progressing MCI
N	401	238	341	212
Age, y	74.45 (5.5)	74.72 (7.9)	72.91 (7.6)	73.61 (7.51)
Education, y	16.36 (2.66)	15.34 (2.9)	16.05 (2.87)	15.82 (2.82)
Sex, % females	49	45	37	39
MMSE	29.1 (1.11)	23.2 (2)	27.91 (1.73)	26.87 (1.74)
ADAS11	5.98 (2.86)	19.85 (6.63)	9.29 (3.9)	13.31 (4.51)
Apoe 4, % zero/one/two alleles	72/26/2	31/48/21	54/36/10	33/51/16
CSF $A\beta_{1-42}$, % positives (no. of subjects with baseline measure)	43 (282)	93 (166)	62 (244)	85 (135)

Positivity to $A\beta_{1-42}$ was defined with respect to the nominal cutoff of 192 pg/mL.

barely missed the selection threshold in component 2 for additional analysis (maximum $P = 0.926$) (Fig. 2).

This study illustrates the potential of effectively combining multivariate statistical modeling in imaging–genetics with recent instruments available from computational biology to lead to insights on the disease pathophysiology. Thanks to the ever-growing data-driven knowledge based on the vast quantities of information now available to the research community, the paradigm proposed in this study may represent a promising avenue for linking imaging–genetics findings to the current knowledge on functional genetics mechanisms involved in neurodegeneration.

Materials and Methods

This section describes the study data, the statistical setting, and the methodology used in this study. Additional details and discussion about the methodological aspects can be found in *SI Materials and Methods*.

Study Participants. Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner. Up-to-date information is available at www.adni-info.org. This research mainly involves further processing of previously collected personal data according to the US ethics regulations. Each subject provided signed informed consent before participation. We have explicit authorization for the use of the ADNI dataset, and we have signed the relevant papers guaranteeing that we abide by the ethics standards.

We selected genotype and phenotype data available in the ADNI-1/GO/2 datasets for 1,192 subjects. Summary sociodemographic, clinical, and genetic information is available in Table 2. At the time of study entry, subjects were diagnosed as healthy individuals ($n = 401$), MCI ($n = 553$), or AD ($n = 238$). A total of 212 (38.3%) MCI patients subsequently converted to AD over the course of the study (6 y). All participants were non-Hispanic Caucasian. AD and MCI groups show significant cognitive decline measured by the Mini Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale (ADAS) Cognitive Subscale (COG) compared with the healthy individuals ($P < 1e{-}2$, two-sample t test for groupwise comparison). There was also a significant increase in individuals with pathological levels of $A\beta_{1-42}$ in the CSF ($A\beta_{1-42} < 192$ pg/mL) across the clinical groups, with proportions ranging from 43% in healthy individuals to 93% in AD patients ($P < 1e{-}2$). Similarly, we observed a higher prevalence of APOE4 carriers in AD and progressing MCI individuals compared with healthy and MCI stable groups. For this analysis, the 639 healthy and AD subjects form the discovery set, while the MCI converters and nonconverters form the independent validation set.

Statistical Discovery. The joint relationship between the genetic and imaging modalities was investigated through PLS modeling (42–44). Among the several PLS versions proposed in the literature, we focus on the symmetric formulation of PLS computed through the singular value decomposition of the cross-covariance matrix (Fig. S1) (43, 44, 45). Within this setting, the aim of PLS is to estimate the latent components that maximize the global covariance between the two input modalities. Each input feature receives a weight in the latent component that represents its relative importance for describing the global joint multimodal relationship. Analyzing these weights helps identify SNPs that are linked to the patterns of cortical thinning in the brain.

In this study, we applied a robust approach for the stable estimation and interpretation of PLS weights in genome-wide genotyping data aimed at promoting sparsity (i.e., selecting only a few features for simplified interpretation) and regularity (by aggregating SNPs within the same genetic neighborhood). This is achieved through a stability selection procedure, in which the reproducibility and robustness of the PLS parameters are assessed through a split-half cross-validation–based scheme on 1 million repetitions of the models on randomly sampled subgroups (Fig. 1 and *SI Materials and Methods*). By considering a predefined partition of each chromosome into contiguous loci of size 10 kb, the procedure leads to the estimation of a confidence measure taking values ranging between 0.0 and 1.0, indicating the probability of each genetic loci containing highly reproducible PLS weights and therefore, serving as a measure of importance of the genomic location (Fig. 2). A similar procedure was used to assess the importance of the phenotype component (Fig. 1). However, no regional binning was used (Fig. 3). The procedure was applied to assess the parameter reproducibility of the first five PLS modes; subsequent analyses were performed only on components with relevant genetic and brain regions (i.e., reproducible PLS weights with selection frequency >95%).

Gene Identification. We analyzed the 10-kb bins (genetic loci) with the selection frequency exceeding 0.95 (i.e., bins selected in 95% or more of the 1 million replications). Within these bins, we then identified the influential SNPs: an SNP was declared influential if it was associated with the weights of greatest magnitude in the PLS components estimated on the full data sample (i.e., SNPs with absolute weights exceeding the 99th quantile of all weights in the component). These weights are the ones contributing to the high selection frequency in the split-half procedure and are representative of the significant variation modeled in the data.

To link SNPs to corresponding genes, we used the computational VEP for GRCh37 with the GENCODE gene annotation. SNPs tagged as "regulatory" were manually investigated and annotated with the nearby genes.

Functional Prioritization. All SNPs successfully annotated with a gene were subjected to functional prioritization through eQTL analysis based on the GTEx data. The sample size in GTEx for relevant brain tissues in AD was rather small (e.g., $n = 81$ for hippocampus). Therefore, we added five more tissues with large sample sizes that were more distantly relevant to AD. Nerve tibial ($n = 256$) was added as a proxy for nervous tissue, whole blood ($n = 338$) and artery tibial ($n = 285$) were included to cover blood-based changes and effects on blood vessels (46), and adipose s.c. ($n = 298$) was selected due to links between AD and obesity, type 2 diabetes, and metabolic disease (47, 48). Finally, transformed fibroblasts ($n = 272$) were included as a general purpose cell line. P values were corrected for multiple testing using the Bonferroni method.

Model Validation in Independent MCI Subjects. The genes that were found to be under expression control by the identified SNPs were validated for their capacity to predict clinical conversion in MCI subjects. To this end, for each identified gene, we applied the PLS weights estimated on the discovery set on the validation set, with the genetic component restricted to SNPs ±20 kb from the gene borders. The identified latent projections (i.e., a weighted sum of SNPs) result in one score per subject per gene. For each gene, the association of the projection score with conversion status was assessed by statistically comparing the scores distribution between healthy individuals and AD patients and between MCI converters and nonconverters (Kruskal–Wallis nonparametric test for two-sample comparison, Bonferroni correction for multiple comparisons).

ACKNOWLEDGMENTS. M.L., J.M.S., D.C.A., and S.O. received support from European Union's Horizon 2020 Research and Innovation Programme Grant 666992 (EuroPOND) for this work. A.A. holds an MRC eMedLab Medical Bioinformatics Career Development Fellowship. This work was supported by Medical Research Council Grant MR/L016311/1. The contribution to this work by B.G. and P.M.T. was funded by NIH "Big Data to Knowledge" Grant U54 EB020403 (principal investigator: P.M.T.). S.W. and C.A. are supported by the National Institute for Health Research (NIHR) Queen Square Biomedical Research Unit in Dementia and Alzheimer's Research UK and the NIHR University College London (UCL) Hospitals Biomedical Research Centre. J.M.S. acknowledges the support of the NIHR UCL Hospitals Biomedical Research Centre; the Wolfson Foundation; Engineering and Physical Sciences Research Council (EPSRC) Grant EP/J020990/1; Medical Research Council (MRC) Grant MR/L023784/1; Alzheimer Research UK (ARUK) Grants ARUK-Network 2012-6-ICE, ARUK-PG2017-1946, and ARUK-PG2017-1946; Brain Research Trust Grant UCC14191; and European Union's Horizon 2020 Research and Innovation Programme Grant 666992. EPSRC Grants EP/J020990/01 and EP/M020533/1 support the work of D.C.A. and S.O. on this topic. S.O. receives funding from EPSRC Grants EP/H046410/1 and EP/K005278, MRC Grant MR/J01107X/1, EU-FP7 Project VPH-DARE@IT Grant FP7-ICT-2011-9-601055, the NIHR Biomedical Research Unit (Dementia) at the UCL, and NIHR University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative-BW. mn.BRC10269). Data collection and sharing for this project were funded by the ADNI (NIH Grant U01 AG024904) and the Department of Defense (DOD) ADNI (DOD Award W81XWH-12-2-0012). The ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering, and it is also funded through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the NIH (https://fnih.org/). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. The ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

1. Rabinovici GD, Jagust WJ (2009) Amyloid imaging in aging and dementia: Testing the amyloid hypothesis in vivo. *Behav Neurol* 21:117–128.
2. Villemagne VL, Okamura N (2014) In vivo tau imaging: Obstacles and progress. *Alzheimers Dement* 10(3 Suppl):S254–S264.
3. Mosconi L, et al. (2010) Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. *J Alzheimers Dis* 20:843–854.
4. Frisoni GB, Fox NC, Jack CR, Jr, Scheltens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* 6:67–77.
5. Bigos KL, Hariri AR, Weinberger DR (2016) *Neuroimaging Genetics: Principles and Practices* (Oxford Univ Press, Oxford).
6. Lambert JC, et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45:1452–1458.
7. Potkin SG, et al. (2009) Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS ONE* 4:e6501.
8. Ramanan VK, et al.; Alzheimer's Disease Neuroimaging Initiative (2014) APOE and BCHE as modulators of cerebral amyloid deposition: A florbetapir PET genome-wide association study. *Mol Psychiatry* 19:351–357.
9. Kam-Thong T, et al. (2012) GLIDE: GPU-based linear regression for detection of epistasis. *Hum Hered* 73:220–236.
10. Stein JL, et al.; Alzheimer's Disease Neuroimaging Initiative (2010) Voxelwise genome-wide association study (vGWAS). *Neuroimage* 53:1160–1174.
11. Schrouff J, et al. (2013) PRoNTo: Pattern recognition for neuroimaging toolbox. *Neuroinformatics* 11:319–337.
12. Szymczak S, et al. (2009) Machine learning in genome-wide association studies. *Genet Epidemiol* 33:S51–S57.
13. Liu J, Calhoun VD (2014) A review of multivariate analyses in imaging genetics. *Front Neuroinform* 8:29.
14. Le Floch E, et al. (2012) Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *Neuroimage* 63:11–24.
15. Vounou M, Nichols TE, Montana G; Alzheimer's Disease Neuroimaging Initiative (2010) Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage* 53:1147–1159.
16. Silver M, Janousova E, Hua X, Thompson PM, Montana G; Alzheimer's Disease Neuroimaging Initiative (2012) Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *Neuroimage* 63:1681–1694.
17. Liu J, et al. (2009) Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum Brain Mapp* 30:241–255.
18. Carithers LJ, et al.; GTEx Consortium (2015) A novel approach to high-quality postmortem tissue procurement: The GTEx project. *Biopreserv Biobank* 13:311–319.
19. Trabzuni D, et al. (2011) Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *J Neurochem* 119:275–282.
20. Krzywinski M, et al. (2009) Circos: An information aesthetic for comparative genomics. *Genome Res* 19:1639–1645.
21. Aken BL, et al. (2016) The Ensembl gene annotation system. *Database (Oxford)* 2016:baw093.
22. Machiela MJ, Chanock SJ (2015) LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31:3555–3557.
23. Du K, Herzig S, Kulkarni RN, Montminy M (2003) TRB3: A tribbles homolog that inhibits Akt/PKB activation by insulin in liver. *Science* 300:1574–1577.
24. Zareen N, Biswas SC, Greene LA (2013) A feed-forward loop involving Trib3, Akt and FoxO mediates death of NGF-deprived neurons. *Cell Death Differ* 20:1719–1730.
25. Aimé P, et al. (2015) Trib3 Is elevated in Parkinson's disease and mediates death in Parkinson's disease models. *J Neurosci* 35:10731–10749.
26. Hua F, et al. (2015) TRB3 links insulin/IGF to tumour promotion by interacting with p62 and impeding autophagic/proteasomal degradations. *Nat Commun* 13:7951.
27. Menzies FM, Fleming A, Rubinsztein DC (2015) Compromised autophagy and neurodegenerative diseases. *Nat Rev Neurosci* 16:345–357.
28. Zhou Y, et al. (2008) E3 ubiquitin ligase SIAH1 mediates ubiquitination and degradation of TRB3. *Cell Signal* 20:942–948.
29. Saleem S, Biswas SC (2017) Tribbles Pseudokinase 3 induces both apoptosis and autophagy in amyloid-β induced neuronal death. *J Biol Chem* 292:2571–2585.
30. Zhang W, et al. (2016) Skeletal muscle TRIB3 mediates glucose toxicity in diabetes and high- fat diet-induced insulin resistance. *Diabetes* 65:2380–2391.
31. Sims-Robinson C, Kim B, Rosko A, Feldman EL (2010) How does diabetes accelerate Alzheimer disease pathology? *Nat Rev Neurol* 6:551–559.
32. Ribe EM, Lovestone S (2016) Insulin signalling in Alzheimer's disease and diabetes: From epidemiology to molecular links. *J Intern Med* 280:430–442.
33. Morris AP, et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 44:981–990.
34. Luciano M, et al. (2011) Whole genome association scan for genetic polymorphisms influencing information processing speed. *Biol Psychol* 86:193–202.
35. Giedraitis V, et al. (2009) Genetic analysis of Alzheimer's disease in the uppsala longitudinal study of adult men. *Dement Geriatr Cogn Disord* 27:59–68.
36. Li H, et al. (2008) Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* 65:45–53.
37. Oláh J, et al. (2011) Interactions of pathological hallmark proteins: Tubulin polymerization promoting protein/p25, β-amyloid, and α-synuclein. *J Biol Chem* 286: 34088–34100.
38. Li MO, Flavell RA (2008) Contextual regulation of inflammation: A duet by transforming growth factor-β and interleukin-10. *Immunity* 28:468–476.
39. Matarin M, et al. (2015) A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. *Cell Rep* 10: 633–644.
40. Guillot-Sestier MV, et al. (2015) Il10 deficiency rebalances innate immunity to mitigate Alzheimer-like pathology. *Neuron* 85:534–548.
41. Kajkowski EM, et al. (2001) β -Amyloid peptide-induced apoptosis regulated by a novel protein containing a g protein activation module. *J Biol Chem* 276: 18748–18756.
42. Wold H (1966) Estimation of principal components and related models by iterative least squares. *Multivariate Analysis* (Academic, New York) pp 391–420.
43. McIntosh AR, Bookstein FL, Haxby JV, Grady CL (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* 3:143–157.
44. Worsley KJ (1997) An overview and some new developments in the statistical analysis of PET and fMRI data. *Hum Brain Mapp* 5:254–258.
45. Friston KJ, Frith CD, Liddle PF, Frackowiak RSJ (1993) Functional connectivity: The principal-component analysis of large (PET) data sets. *J Cereb Blood Flow Metab* 13: 5–14.
46. Kimbrough IF, Robel S, Roberson ED, Sontheimer H (2015) Vascular amyloidosis impairs the gliovascular unit in a mouse model of Alzheimer's disease. *Brain* 138: 3716–3733.
47. Luchsinger JA, Gustafson DR (2009) Adiposity, type 2 diabetes, and Alzheimer's disease. *J Alzheimers Dis* 4:693–704.
48. Ferreira ST, Clarke JR, Bomfim TR, De Felice FG (2014) Inflammation, defective insulin signaling, and neuronal dysfunction in Alzheimer's disease. *Alzheimers Dement* 10 (1 Suppl):S76–S83.

SCIENTIFIC REPORTS

OPEN

Multimodal Image Analysis in Alzheimer's Disease via Statistical Modelling of Non-local Intensity Correlations

Received: 29 June 2015
Accepted: 04 February 2016
Published: 11 April 2016

Marco Lorenzi[1], Ivor J. Simpson[1], Alex F. Mendelson[1], Sjoerd B. Vos[1,2], M. Jorge Cardoso[1], Marc Modat[1], Jonathan M. Schott[3] & Sebastien Ourselin[1]

The joint analysis of brain atrophy measured with magnetic resonance imaging (MRI) and hypometabolism measured with positron emission tomography with fluorodeoxyglucose (FDG-PET) is of primary importance in developing models of pathological changes in Alzheimer's disease (AD). Most of the current multimodal analyses in AD assume a *local* (spatially overlapping) relationship between MR and FDG-PET intensities. However, it is well known that atrophy and hypometabolism are prominent in different anatomical areas. The aim of this work is to describe the relationship between atrophy and hypometabolism by means of a data-driven statistical model of non-overlapping intensity correlations. For this purpose, FDG-PET and MRI signals are jointly analyzed through a computationally tractable formulation of partial least squares regression (PLSR). The PLSR model is estimated and validated on a large clinical cohort of 1049 individuals from the ADNI dataset. Results show that the proposed non-local analysis outperforms classical local approaches in terms of predictive accuracy while providing a plausible description of disease dynamics: early AD is characterised by non-overlapping temporal atrophy and temporo-parietal hypometabolism, while the later disease stages show overlapping brain atrophy and hypometabolism spread in temporal, parietal and cortical areas.

The multimodal analysis of anatomical and physiological images is of primary importance in developing comprehensive models of biological processes and pathologies, and increasing the statistical power of current imaging biomarkers. Already, both brain atrophy, measured in magnetic resonance images (MRIs), and hypometabolism, quantified by positron emission tomography with fluorodeoxyglucose radiotracers (FDG-PET), are among the primary diagnostic biomarkers of Alzheimer's disease (AD). The information provided by these two imaging modalities is correlated, since hypometabolism and neuronal loss are interdependent biological phenomena. However, at the present moment, a joint model of the hypometabolism-atrophy relationship in AD has not been developed, and current hypotheses on their interaction are mostly based on the quantification of grey matter volume and FDG uptake at the regional level.

In recent years, voxel-wise approaches to multimodal analysis in AD have been proposed[1]. In particular, image synthesis techniques based on machine learning have been used to synthesise FDG-PET images from MRIs of AD patients for diagnostic purposes[2,3]. The majority of these approaches are based on the local modelling of the relationship between MR and FDG-PET signals, either by considering the voxels independently, or through neighborhoods (patches) defined around voxels. However, it is well known that the link between morphology and function in the brain is not purely local[4]. For this reason, local methods may provide only a limited description of the link between structure and function in AD.

Several techniques have been proposed for modelling non-overlapping signal correlations in the field of functional MRI analysis. For instance, both independent component analysis (ICA) or partial least square (PLS) approaches have been successfully applied to the joint analysis of functional activation in the brain and covariates drawn from genetic, clinical, or imaging data[5–7]. In the context of correlation modelling in multimodal imaging analysis, multivariate techniques such as PLS have the appealing characteristic that they do not rely on any

[1]Translational Imaging Group, CMIC, UCL, London, UK. [2]MRI Unit, Epilepsy Society, Chalfont St Peter, UK. [3]Dementia Research Centre, UCL Institute of Neurology, Queen Square, London, UK. Correspondence and requests for materials should be addressed to M.L. (email: m.lorenzi@ucl.ac.uk)

hypothesis about the spatial overlap between voxels' signals. They are thus able to model relationships between non-adjacent voxels. Unlike purely local correlation model approaches (for example, those based on voxel-wise correspondencies or on patch-based search windows), these methods optimise the *latent components* describing the *global* correlation of the images treated as multidimensional arrays. This enables them to model the potentially significant interactions between voxels located in completely different areas of a single image, or between voxels in images of different modalities.

Several multivariate approaches have previously been applied to the multimodal analysis of imaging data in neurodegenerative diseases[8]. Notable approaches include parallel ICA, which has been used to analyze the relationships between brain amyloid deposition and either atrophy or hypometabolism[9,10], and canonical correlation analysis, which has been used to study the correlation between structural connectivity and brain atrophy[11]. Though PLS itself has been previously applied in the joint analysis of brain metabolism and atrophy[12,13], past analyses have been limited to relatively small clinical samples and have focused on solely the first latent component. The use of higher-order components may aid in the discovery of more complex correlation structures, though it brings with it greater challenges related to stability and replicability.

The aim of this work is to investigate the spatial relationship between brain atrophy and hypometabolism in a large clinical cohort of the ADNI dataset, by means of a data-driven PLS statistical model of non-overlapping intensity correlations. This is achieved by applying a computationally tractable formulation of PLS regression (PLSR) to the joint analysis of non-local intensity correlations in FDG-PET and T1 weighted MR images. Unlike previous studies, in this work we extend PLSR to the analysis of the high-order latent components, and we introduce a thorough cross-validation scheme in order to identify the reproducible and biologically relevant latent components of joint correlation.

The performance of the PLSR model is compared to a common non-parametric approach for multimodal image analysis based on local intensity similarities. The experimental validation shows that the proposed PLSR approach outperforms the local reference analysis in terms of predictive accuracy while providing an interpretable, reproducible and biologically plausible description of the spatial relationship between atrophy and hypometabolism in AD.

Local vs Non-local Correlation Models of Imaging Data

In this section, we introduce the computational models used to compare non-local and local assumptions of multimodal intensity correlation. The respective models are a computationally tractable application of PLSR to non-local intensity correlation in image data[14] and a non-parametric model based on local patch similarities[15,16].

Computationally Tractable PLSR in Imaging Data. In the following, let $\mathbf{X} = (X_k)_{k=1}^{N_s}$ and $\mathbf{Y} = (Y_k)_{k=1}^{N_s}$ be the matrices of predictor and predicted image modalities respectively, where $\{X_k,\ Y_k\}$ is the multimodal image pair sampled at the same voxel grid, of subject k. We assume that $\mathbf{X}$ and $\mathbf{Y}$ correspond to T1-MR and FDG-PET respectively. The size of $\mathbf{X}$ and $\mathbf{Y}$ is $N_s \times N$, where N_s is the number of individuals, N is the number of image voxels, and the images are represented by row vectors.

The partial least squares (PLS) approach is based on the decomposition of the observations through a projection onto m-dimensional latent spaces defined by the basis vectors $\mathbf{T} = (\mathbf{t}_l)_{l=1}^{m}$ and $\mathbf{U} = (\mathbf{u}_l)_{l=1}^{m}$ such that $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$, and $\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}$, where $\mathbf{P}$ and $\mathbf{Q}$ are the associated coefficients, and $\mathbf{E}$ and $\mathbf{F}$ are matrices of residuals. In particular, PLS aims to maximise the covariance of the projections in the latent space: $cov(\mathbf{t}, \mathbf{u})^2 = max_{\mathbf{w},\mathbf{c}}[cov(\mathbf{Xw}, \mathbf{Yc})]^2$, where $\mathbf{w}$ and $\mathbf{c}$ are unitary basis vectors of the latent space. Several formulations of PLS have been proposed in different research contexts[17–21], and it can be shown that the solution of PLS can be obtained from the principal vectors of the singular value decomposition (SVD) of the covariance matrix $\mathbf{X}^T\mathbf{Y}$[18]. PLS can be iteratively computed as follows:

Let $\mathbf{X}^{(0)} = \mathbf{X}$, $\mathbf{Y}^{(0)} = \mathbf{Y}$. Iterate over the index $i = 1, \ldots, m \leq N_s$:

1. *SVD step.* Compute the principal eigen-vectors $\mathbf{w}_i$ and $\mathbf{c}_i$ of the SVD decomposition of the matrix $\mathbf{R}_i = \mathbf{X}^{(i)T}\mathbf{Y}^{(i)}$.
2. Compute the latent vectors $\mathbf{t}_i = \mathbf{X}^{(i)}\mathbf{w}_i$, and $\mathbf{u}_i = \mathbf{Y}^{(i)}\mathbf{c}_i$, and corresponding coefficients $\mathbf{p}_i = \mathbf{X}^{(i)T}\mathbf{t}_i$.
3. *Deflation step.* Decorrelate the data from the principal directions: $\mathbf{X}^{(i+1)} = \mathbf{X}^{(i)} - \mathbf{t}_i\mathbf{p}_i^T$, and $\mathbf{Y}^{(i+1)} = \mathbf{Y}^{(i)} - \mathbf{u}_i\mathbf{c}_i^T$.

We note that the matrix $\mathbf{R}_i$ is usually very large (voxels × voxels), and its SVD decomposition is generally computationally infeasible. However, the SVD step can still be efficiently computed from the eigen-value problem associated with the matrix $\mathbf{XX}^T\mathbf{YY}^T$, which is usually of much smaller dimension ($N_s \times N_s$). This approach has been proposed previously in[14], which focused on the analysis of within-modality non-local intensity correlations in neuroimages. In particular, this efficient optimization scheme was used to model either group-wise patterns of cortical thickness from MRI, or functional connectivity networks measured in fMRI. In this work, we apply this computational approach in the context of multimodal analysis of brain images.

PLSR builds upon the above formulation of PLS by assuming a linear relationship between the vectors $\mathbf{t}$ and $\mathbf{u}$, i.e. $\mathbf{U} = \mathbf{TB} + \mathbf{H}$, where $\mathbf{B}$ is a latent linear mapping and $\mathbf{H}$ is the residual matrix. The PLS model can thus be rewritten as $\mathbf{Y} = \mathbf{TC}^T + \mathbf{F}*$, where $\mathbf{C}^T = \mathbf{BQ}^T$, and $\mathbf{F}* = \mathbf{HQ}^T + \mathbf{F}$ is the residual error. It can be shown that the solution of the PLSR is $\widetilde{Y} = X\mathbf{D}_{\mathbf{PLS}}$ where the regression coefficient $\mathbf{D}_{PLS} = (\mathbf{P}^T)^+\mathbf{BQ}^T$, and $(\mathbf{P}^T)^+ = \mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}$ is the right Moore-Penrose pseudoinverse of $\mathbf{P}^T$[20].

	healthy	MCI stable	MCI conv	AD
N	274	453	154	168
age (years)	74.1 (5.98)	72.16 (7.55)	73.21 (7.37)	75.66 (7.66)
sex (% females)	48	42	41	42
education (years)	16.22 (2.77)	16.01 (2.73)	16.03 (2.67)	15.1 (3.08)
MMSE	28.98 (1.21)	28.11 (1.66)	27.02 (1.75)	23.05 (2.1)

Table 1. Baseline socio-demographical and clinical information of the cohort of this study. The entries for age, education and MMSE indicate group-wise mean and standard deviation in parenthesis. MMSE: mini mental state exam.

PLSR model of the non-overlapping spatial correlations in multimodal images. PLSR has a number of appealing features that can be exploited in the context of high-dimensional modelling of multimodal images. First, the basis of spatial eigen-components, $\{\mathbf{w}_i, \mathbf{c}_i\}$, provides a *parsimonious* and low-dimensional representation of the multimodal correlation pattern, and can be used for exploratory analysis and modelling purposes. Second, the loadings $(X^T\mathbf{w}_i)_{i=1}^m$ are a low-dimensional representation of the individual anatomy that can be used to address quantitative analysis problems, such as group-wise comparisons or classification. Finally, PLSR defines a transfer function linking the two modalities and, given an individual image X, provides a prediction of the associated target modality Y according to the model of non-local signal correlation estimated in the data $(\mathbf{X}, \mathbf{Y})$.

Relationship between PLSR and principal component regression. Another intuitive approach to multimodal correlation modelling in image data is principal component regression (PCR). This consists of an initial PCA step applied separately to the predicted and predictor variables, and a subsequent regression step to model the correlations between the resulting low dimensional representations. Both PCR and PLSR yield a predictive model using low dimensional latent space, but they differ in how this is driven; while PLSR aims to find a subspace that directly maximises the covariance between the predicted and predictor variables, PCR selects a latent space that maximises the variability within each variable set separately. For this reason, it may include components that are not useful in characterising the relationship between variable sets which may degrade predictive performance. We therefore prefer PLSR. The interested reader can find an experimental illustration of the differences between PLSR and PCR when applied to the problem analysed in this study in Appendix A.

Local Models of Intensity Correlations Based on Patch Similarities. Patch-based methods are becoming a popular approach for the estimation of non-linear signal correspondences between different modalities. They have found several successful applications in medical image analysis, such as multi-modal image registration[15], or in FDG-PET synthesis[2]. This approach is based on the assumption that, given an individual image X, the intensities of the target modality $\widetilde{Y}$ can be inferred from the local intensity correspondences observed in a database of atlas pairs $\{X_j^A, Y_j^A\}_{j=1}^{N_A}$[22]. The between-modality voxel-to-voxel mapping $X(s) \rightarrow \widetilde{Y}(s)$ is usually not explicitly modelled in a parametric way, but is indirectly inferred from the local intensities of the images $Y_j^A(s)$ corresponding to the atlases X_j^A most correlated with X. The local correlation model presented here is the same as the one proposed in state-of-art approaches in the context of FDG-PET image synthesis[16]. All subjects images are aligned to the target using non-linear registration, and the intensity at a given location is estimated using the most similar patch in the database as determined using the local intensity information (the L^2 metric). The chosen patch size was of 5 voxels.

Analysis of Brain Hypometabolism and Atrophy in Alzheimer's Disease

Study Participants. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu, Date of Access: 04/2013). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease.

Patients were selected when both MR and FDG-PET images were available at the baseline timepoint. The resulting study cohort consisted of 1049 subjects: 274 healthy controls, 168 patients affected by AD, and 607 patients with mild cognitive impairment (MCI). Of the latter, 154 subsequently converted to AD during the time of the study. Clinical and socio-demographical information are reported in Table 1.

Image Processing. FDG-PET images were obtained at the standardised resolution of 8 mm FWHM, and normalised using the mean intensity in the cerebellar grey matter. T1-weighted MR images at both 1.5 and 3.0 Tesla were included to increase the size of the available sample. A sample specific group-wise space was defined for our analysis using iterative non-rigid registration and averaging the grey matter segmented from the MR images. Registration was performed using the freely available nifty-reg package[23], and grey matter and FDG-PET images were resampled to the group-wise space. The resampled grey matter images were modulated by the corresponding Jacobian determinant of the template-to-subject transformation, and subsequently spatially filtered at the point spread function of the PET images and downsampled. Thanks to the modulation and to the downsampling

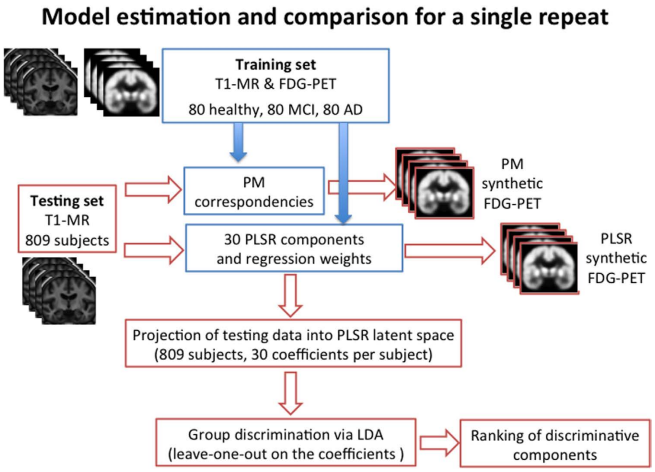


Figure 1. Flowchart of the cross-validation scheme of a single repeat in the proposed experimental setting.

operation, the resulting anatomical areas of apparent ventricular and CSF expansion are associated to smaller voxel-wise intensity values due to the scaling by the Jacobian determinant values.

Statistical Analysis. *Model Estimation and Comparison.* The goodness of fit of PLSR and local approaches was assessed by cross-validation. The training data was composed of 80 healthy controls, 80 MCI, and 80 AD patients, randomly chosen from the study cohort. The data were respectively used to 1) estimate the PLSR latent components and regression coefficients, and 2) as an atlas database for the local patch-based method (PM). The PLSR model was computed by estimating 30 latent components. The model could be ideally computed by estimating the 239 latent basis-components corresponding to $N - 1$ training data samples. The number of estimated PLSR components was limited to 30 for practical reasons because, as shown in the next experimental section, the stability and reproducibility of the high-order components is generally very low, and they usually provide very little contribution to the model performance. The resulting multimodal correlation models were validated on the remaining subjects. The experiment was repeated 10 times with different training sets to ensure the generalisation of the results. Due to the non-parametric nature of PM, we also compared the PM trained with a leave-one-out scheme (1048 training samples per test) in order to use of the largest amount of training data. The predictive accuracy was measured by the absolute difference between predicted and observed FDG-PET in temporal, posterior and parietal cortices, and by comparing the average predicted FDG-PET regional values to the SUVR values independently reported in the ADNI dataset. Figure 1 illustrates the flowchart of the cross-validation scheme of a single repeat adopted in the proposed experimental setting.

Reproducibility and Biological Plausibility of PLSR model. We investigated the biological plausibility of the non-local correlation pattern of the PLSR model. For this purpose, a linear discriminant analysis (LDA) was performed on the coefficients of the latent space associated with the testing subjects, in order to identify the mostly discriminative PLS components through leave-one-out. The discriminative accuracy of the PLSR model was quantified by computing the area under the receiver operating characteristic (ROC) curve associated to the LDA classification result.

The interpretation of the PLSR modes of correlation is usually challenging, since some of the obtained components (especially at the high order) tend to be noisy and not necessarily related to meaningful anatomical interpretation (an illustrative example of the set of the first 10 components estimated by PLSR in a single repeat is shown in Appendix B). For this purpose, in order to address this important issue of robustness, we measured the reproducibility of the most discriminative components across the 10 repeats. We were interested primarily in the rate of reproduction of the individual components, rather than in their relative ordering. To this end, the discriminative power of each component was quantified by the absolute value of the associated LDA weight, and the resulting 5 mostly discriminative components of each experiment were matched to those of the other repeats. Components were matched when the absolute value of the correlation between them exceeded 0.5. When multiple matches were possible, the strongest was chosen. The resulting labelling establishes the reproducibility of the discriminative components across repeats, and does not necessarily reflect the order of the eigen-components estimated in each PLS run.

Results

Model comparison. Figure 2 shows the average pattern of absolute differences of the predictions in AD and healthy controls with both methods. We notice that PLSR generally provides a better fit, while the local patch based method leads to larger estimation errors in parietal and temporal areas. We note that increasing the training sample size slightly improves the PM, especially in the temporal regions. The average regional absolute error between predicted images and real ones was systematically higher for PM as compared to PLSR, and significantly different for 8 out of 10 repeats ($p < 0.01$, paired t-test). The PLSR prediction also provided significantly better agreement with the ADNI measurements than PM (Table 2A). This is reflected by the significantly higher effect

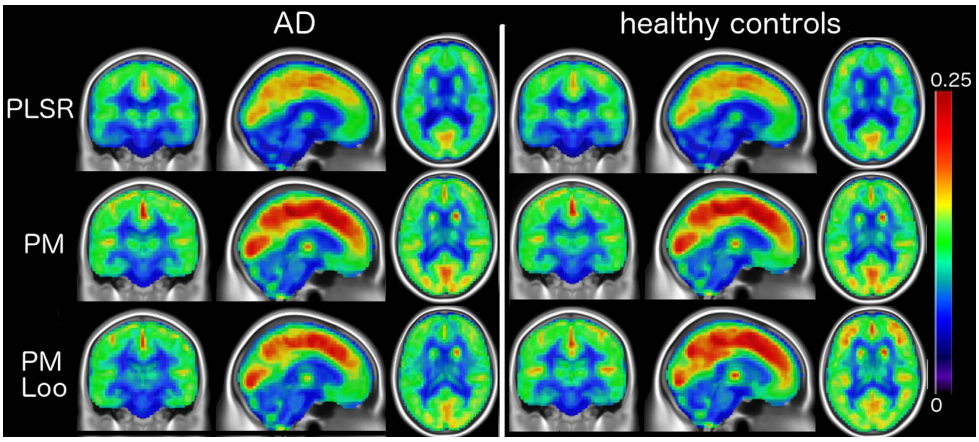


Figure 2. Mean absolute prediction error of PLSR and PM (240 training samples, and leave-one-out -Loo-). PLSR provides higher predictive accuracy than the local patch based (PM) approach. Results are similar when considering the MCI group (not shown).

A. Correlation wrt ADNI SUVR			B. Effect size		
	PLSR	PM		PLSR	PM
Whole cohort*	0.31 (0.25,0.38)	0.21 (0.14, 0.27)	AD vs HC *	1.07	0.79
AD and HC*	0.33 (0.22, 0.43)	0.23 (0.12, 0.34)	MCIc vs MCIs	0.53	0.37
MCIc and MCIs*	0.30 (0.22, 0.38)	0.20 (0.12, 0.28)	MCIc vs HC*	0.67	0.46

Table 2. A. Correlation (mean, 95% confidence interval) between predicted average regional FDG-PET and the corresponding SUVR values reported in ADNI. B. Effect size between the measures obtained with PLSR and with PM. HC: healthy controls, MCIc: MCI converted to AD, MCIs: MCI stable. (* for significant differences, $p < 0.05$, paired t-test).

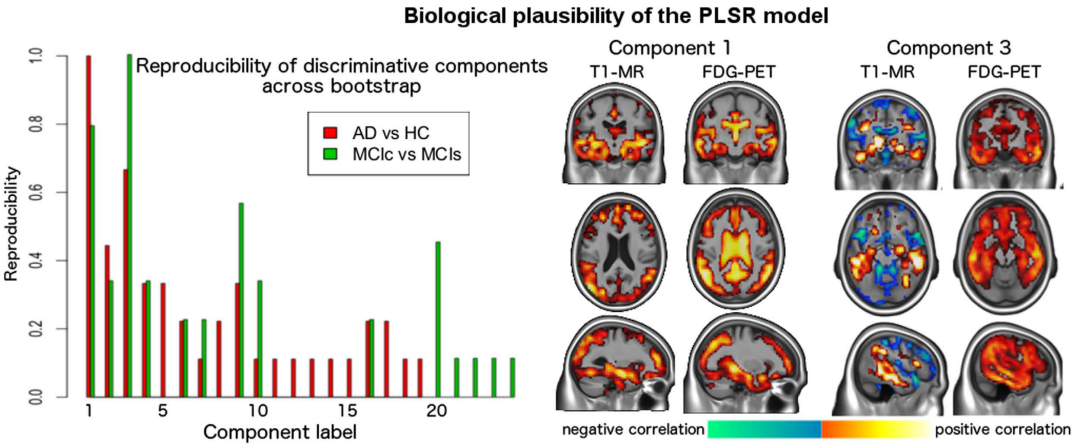


Figure 3. Left: reproducibility of the discriminative components. When comparing AD and controls, component 1 was the only one 100% reproducible and discriminative across repeats. The same consideration applies to component 3 when comparing stable and converting MCI. Right: component 1 describes the relationship between atrophy and FDG-PET uptake spread in temporal, parietal and posterior regions. We also note the partial volume effect in the ventricles for the FDG component. Component 3 shows the *non-overlapping* spatial inverse relationship between increased expansion of the CSF (ventricles and brain sulci), and joint increased temporal atrophy and cortical hypometabolism.

size associated with the average measures for the PLSR approach, indicating a better separation between clinical groups (Table 2B). 212

Biological Plausibility of the PLSR model. Figure 3, left, shows the reproducibility results for the PLSR components across repeats. The components are ordered according to the order of output in the PLSR results obtained in the first cross-validation repeat. The reproducibility results are instead quantified by the green and red bars, which indicates the number of times that a component was among the most discriminative across repeats.

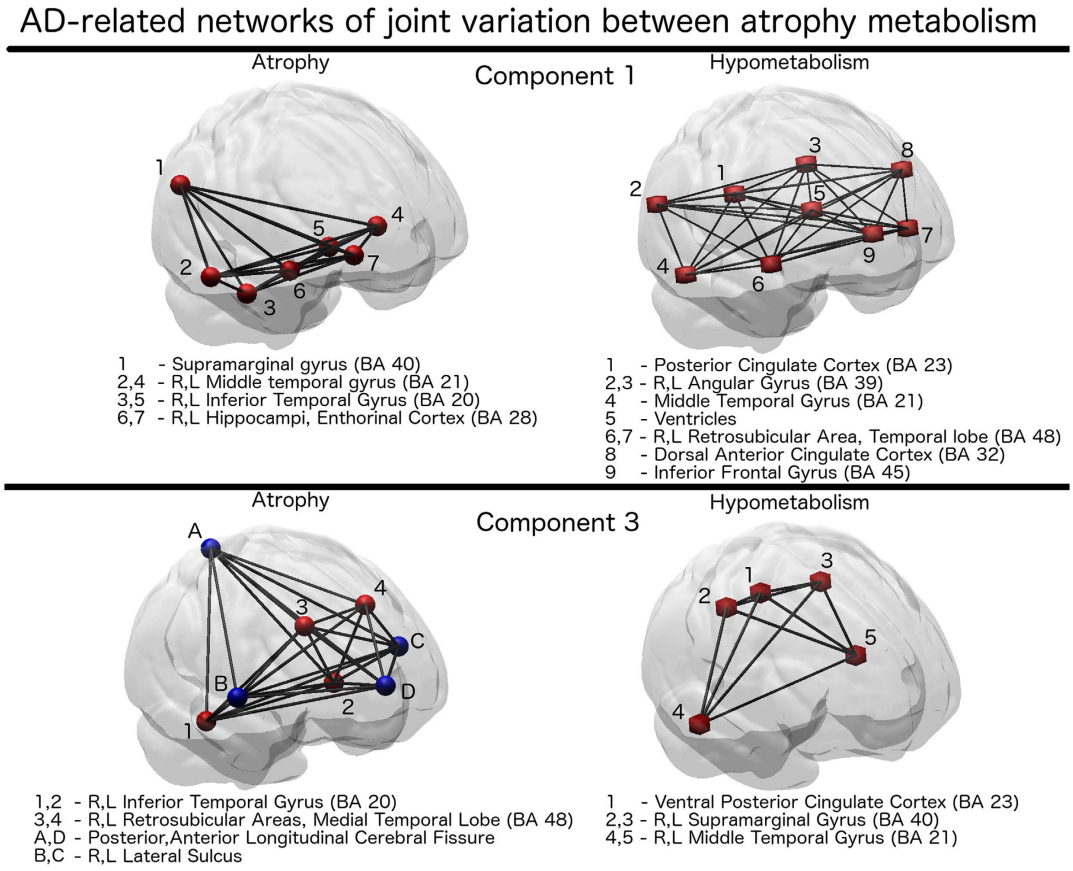


Figure 4. PLSR networks of joint relationship between atrophy and hypometabolism derived from Component 1 and Component 3. Red hubs indicate regions of joint within- and between-modality correlation. Blue hubs indicate anticorrelated regions (CSF expansion measured in T1-MR). BA = Broadmann anatomical areas.

We note that only few of the mostly discriminative PLSR components are highly reproducible across repeats. In particular, the first eigen-component estimated in each PLSR repeat is the most discriminative when comparing AD vs healthy controls, while the third one is most discriminative when comparing both MCI stable and converters, and healthy controls and MCI converters. Both components were 100% reproducible across repeats. We note that since the figure is relative to the first cross validation repeat, the components estimated during the other repeats that did not find any match in the reproducibility analysis are omitted in the figure. However, apart from the reported components 1 and 3, no other reproducible component associated to the other repeats was observed in the analysis.

These components are shown on the right hand side of Fig. 3, and the associated correlation network is shown in Fig. 4. A 3D rendering of the correlation networks is shown at the following url: https://www.dropbox.com/s/orsf3nt6hq2kp38/supplementary_animations1.mov?dl=0 (12/10/2015). The networks were obtained by thresholding components 1 and 3, and by subsequently applying a morphological opening operation in order to identify a consistent set of clusters of maximal PLSR weights.

On one hand, we note that component 1 describes the relationship between atrophy and FDG-PET uptake spread in temporal, parietal and posterior regions. In particular, it shows the partial volume effect due to ventricular expansion in AD, that is already observable in the raw data, and that leads to the very large variation of the FDG-PET signal in the ventricles in subjects with pronounced global brain atrophy. On the other hand, component 3 shows the non-overlapping inverse correlation pattern between increased expansion of the CSF, and joint increased temporal atrophy and cortical hypometabolism.

Finally, the average area under the ROC curve for the classification tasks across the different folds was 0.87 (0.83, 0.91–95% c.i.) for the comparison of AD vs healthy controls, and 0.75 (0.73, 0.76–95% c.i.) for the comparison between stable and converting MCI. This result primarily confirms the ability of the proposed PLSR to model biologically relevant features, and is in line with the classification performance based on T1-MR information previously reported in the literature on the ADNI dataset[24–26].

Conclusions

We have investigated the problem of multimodal analysis of biomedical images in AD, by comparing two different modelling hypothesis based on state-of-art techniques, PLSR and patch-based local correlation, to promote non-local correlation analysis approaches with respect to localized ones in describing multimodal correlation

patterns in AD. Our study introduces and validates the use of PLSR in the context of multimodal modelling in AD by showing that PLSR bases and coefficients can be estimated in very large datasets of volumetric images through a computationally tractable approach to the eigen-decomposition.

Non-local vs local multimodal modelling in AD. Our results show that the proposed non-local approach outperforms classical PM-based multimodal local correlation models in terms of modelling accuracy and predictive power. The ensemble of the reported results proves the ability of the proposed PLSR in capturing biologically relevant features, and in generalising to unseen structural imaging data of T1-MR scans.

Even though the presented study does not provide a theoretical proof of the superiority of non-local methods, our results show that T1-MR and FDG-PET present reproducible and consistent patterns of correlations between non-overlapping anatomical areas. This study thus shows that realistic multimodal models of neurodegeneration necessarily need to account for the non-local relations intimately related to the neurobiological aspects of the disease.

Plausibility of the PLSR model. PLSR provides a parsimonious description of the global biological variability, represented by the low-dimensional latent subspace parameterisation. For this reason, the interpretation and statistical analysis of PLSR is more straightforward than that of the usually complex models provided by non-parametric local approaches[2].

Our analysis revealed that in the sequential stages of the pathology (cognitively normal $->$MCIc $->$AD) we can consistently identify two reproducible components of correlation between atrophy and hypometabolism. Our results are therefore supportive of the existence of different patterns of atrophy and hypometabolism which differentially characterise the different stages of the disease, and thus are informative of the dynamics of the pathology.

The correlation networks highlighted in this study are supported by known biological dynamics between atrophy and hypometabolism in dementia[4]: although hypometabolism and atrophy are typically locally correlated, i.e. areas with neuronal loss (atrophy) show by definition reduced or absent metabolism, hypometabolism may be seen in areas not obviously or typically affected by atrophy, as exemplified by focal dementia syndromes, such as posterior cortical atrophy.

This work shows that T1 weighted MRI and FDG-PET in AD are highly correlated and share important patterns of common non-overlapping spatial relationship. The proposed method could be used in the future to identify and decorrelate the common inter-modality variation from biomedical images for the identification of more specific image based biomarkers.

References

1. Frisoni, G. B. *et al. In vivo* mapping of amyloid toxicity in Alzheimer disease. *Neurology* **72(17),** 1504–11 (2009).
2. Rongjian, L. *et al.* Deep learning based imaging data completion for improved brain disease diagnosis. *Proc. of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS Springer, **8675**, 305–312 (2014).
3. Burgos, N. *et al.* Subject-specific Models for the Analysis of Pathological FDG PET Data. *Proc. of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS Springer, **9351**, 651–658 (2015).
4. Jack, C. R. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol* **9(1),** 119–128 (2010).
5. Krishnan, A. *et al.* Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage*, **56(2),** 455–475 (2011).
6. Vergara, V. M. *et al.* A three-way parallel ICA approach to analyze links among genetics, brain structure and brain function. *NeuroImage* **98,** 386–394 (2014).
7. Burzynska, A. Z. *et al.* A scaffold for efficiency in the human brain, *J Neurosci* **33(43),** 17150–17159 (2013).
8. Sui, J., Adalib, T., Yu, Q. & Calhoun V.D. A review of multivariate methods for multimodal fusion of brain imaging data. *J Neurosci. Methods* **204(1),** 68–81 (2012).
9. Tosun, D. *et al.* Spatial patterns of brain amyloid-β burden and atrophy rate associations in mild cognitive impairment, *Brain* **134(4),** 1077âĂŞ1088 (2011).
10. Laforce, R. *et al.* Parallel ICA of FDG-PET and PiB-PET in three conditions with underlying Alzheimer's pathology. *NeuroImage: Clinical* **4,** 508–516.
11. Avants, B. *et al.* Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage*. **50(3),** 1004–1016 (2010).
12. Chen, K. *et al.* Linking functional and structural brain images with multivariate network analyses: a novel application of the partial least square method. *NeuroImage* **47(2),** 602–610 (2009).
13. Chen, K. *et al.* Correlations between FDG PET glucose uptake-MRI gray matter volume scores and apolipoprotein E ε4 gene dose in cognitively normal adults: a cross-validation study using voxel-based multi-modal partial least squares. *NeuroImage* **60(4),** 2316–2322 (2012).
14. Worsley, K. J., Chen, J., Lerch, J. & Evans, A. C. Comparing Functional Connectivity via Thresholding Correlations and Singular Value Decomposition. *Phil Trans: Biological Sciences* **360(1457),** 913–920 (2005).
15. Iglesias, J. E. *et al.* Is Synthesizing MRI Contrast Useful for Inter-modality Analysis? *Proc. of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS Springer, **8149**, 631–638 (2013).
16. Burgos, N. *et al.* Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Application to Brain Studies. *IEEE TMI* **33(12),** 2232–42 (2014).
17. Wold, H. *et al.* Estimation of principal components and related models by iterative least squares. *Multivariate analysis* **1,** 391–420 (1966).
18. Agnar, H. PLS regression methods. *J Chemometrics* 2, 3 John Wiley & Sons, Ltd. (1988).
19. Geladi, P. & Kowalski, B. R. Partial least-squares regression: a tutorial. *Analytica chimica acta* **185,** 1–17 (1986).
20. Rosipal, R. & Krämer, N. Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection Techniques*. LNCS, Springer, 34–51 (2006)214
21. Vinzi, V. *et al. Handbook of partial least squares* Springer, 1–22 (2010).
22. Hertzmann, A. *et al. Image analogies*, Proceedings of the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH), Los Angeles, 327–340. New York: ACM (2001).
23. Modat, M. *et al.* Fast free-form deformation using graphics processing units. *Comput Meth Prog Bio* **98(3),** 278–284. http://sourceforge.net/projects/niftyreg (18/09/2009) (Date of access: 03/2014).

24. Chincarini, A. *et al.* Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. *NeuroImage*, **58(2),** 469–80 (2011).
25. Young, J. *et al.* Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical* **2,** 735–745 (2013).
26. Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. M. & Trojanowski, J. Q. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* **32(12),** 2322.e19–2322.e27 (2011).

Acknowledgements

Sebastien Ourselin receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative-BW.mn.BRC10269). Jonathan M. Schott acknowledges the support of the NIHR UCL/H Biomedical Research Centre, the NIHR Queen Square Biomedical Research Unit in Dementia, and grant funding from the MRC, Alzheimer's Research UK, and Leonard Wolfson Experimental Neurology Centre. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

Author Contributions

M.L. designed and implemented the PLSR code, performed the statistical analysis and wrote the main manuscript text. I.J.S. and M.J.C. provided feedback on the statistical content, and helped in the designing of the local correlation technique of section 2.3. A.F.M. and M.M. worked on the preprocessing of the ADNI data used in the experimental section 4.1. S.B.V. provided support in the conception and visualization of the networks of Figure 4. J.S. and S.O. supervised the project, and provided feedback on the analysis and on the interpretation of the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Lorenzi, M. *et al.* Multimodal Image Analysis in Alzheimer's Disease via Statistical Modelling of Non-local Intensity Correlations. *Sci. Rep.* **6**, 22161; doi: 10.1038/srep22161 (2016).

Selected publications for Axis III

Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data

Santiago Silva[1], Boris Gutman[2], Eduardo Romero[3], Paul M. Thompson[4], Andre Altmann[4] and Marco Lorenzi[1]

1. Université Côte d'Azur, INRIA Sophia Antipolis, EPIONE research group, France

[2] Department of Biomedical Engineering, Illinois Institute of Technology, USA

[3] CIM@LAB, Universidad Nacional de Colombia, Bogotá, Colombia

[4] USC Stevens Institute for Neuroimaging and Informatics, Los Angeles, USA

[5] Centre for Medical Image Computing, UCL, London, UK

Originally published in:
IEEE International Symposium on Biomedical Imaging (ISBI), Venice, 2019

Abstract

At this moment, databanks worldwide contain brain images of previously unimaginable numbers. Combined with developments in data science, these massive data provide the potential to better understand the genetic underpinnings of brain diseases. However, different datasets, which are stored at different institutions, cannot always be shared directly due to privacy and legal concerns, thus limiting the full exploitation of big data in the study of brain disorders. Here we propose a *federated learning* framework for securely accessing and meta-analyzing any biomedical data without sharing individual information. We illustrate our framework by investigating brain structural relationships across diseases and clinical cohorts. The framework is first tested on synthetic data and then applied to multi-centric, multi-database studies including ADNI, PPMI, MIRIAD and UK Biobank, showing the potential of the approach for further applications in distributed analysis of multi-centric cohorts.

1 Introduction

Nowadays, a large amount of magnetic resonance images (MRI) scans are stored across a vast number of clinical centers and institutions. Researchers are currently analyzing these large datasets to understand the underpinnings of brain diseases. However, due to privacy concerns and legal complexities, data hosted in different centers cannot always be directly shared. In practice, data sharing is also hampered by the need to transfer large volumes of biomedical data with the associated bureaucratic burden. This situation led researchers to look for an analysis solution within *meta-analysis* or *federated learning* paradigms. In the federated setting, a model is fitted without sharing individual information across centres, but only model parameters. Meta-analysis instead performs statistical testing by combining results from several independent assays[1], for example by sharing p-values, effect sizes, and/or standard errors across centers.

One example of such a research approach is the *Enhancing NeuroImaging Genetics through Meta-Analysis* (ENIGMA) consortium (enigma.usc.edu). With a large number of institutions worldwide [2], ENIGMA has become one of the largest networks bringing together multiple groups analyzing neuroimaging data from over 10,000 subjects. However, most of ENIGMA's secure meta-analytic studies in neuroimaging are performed using mass-univariate models.

The main drawback of mass-univariate analysis is that they can only model a single dependent variable at a time. This is a limiting assumption in most of the biomedical scenarios (e.g., neighboring voxels or genetic variations are highly correlated). To overcome this problem, multivariate analysis methods have been proposed to better account for covariance in high-dimensional data.

In a federated analysis context, a few works proposed generalization of standard neuroimaging multivariate analysis methods, such as Independent Component Analysis [3], sparse regression, and parametric statistical testing [4, 5]. Since these methods are mostly based on stochastic gradient descent, a large-number of communications across centers may be required to reach convergence. Therefore, there is a risk of computational and practical bottlenecks when applied to multi-centric high-dimensional data.

Lorenzi *et* al.[6, 7] proposed a multivariate dimensionality reduction approach based on eigenvalue decomposition. This approach does not require iteration over centers, and was demonstrated on the analysis of the joint variability in imaging-genetics data. However, this framework is still of limited practical utility in real applications, as data harmonization (e.g., standardization and covariate adjustment) should be also consistently performed in a federated way.

Herein we contribute to the state-of-the-art in federated analysis of neuroimaging data by proposing an end-to-end framework for data standardization, confounding factors correction, and multivariate analysis of variability of high-dimensional features. To avoid the potential bottlenecks of gradient-based optimization, the framework is based on schemes analysis through *Alternating Direction Method of Multipliers* (ADMM) reducing the amount of iterations.

We illustrate the framework leveraging on the ENIMGA Shape tool, to provide a first application of federated analysis compatible with the standard ENIGMA pipelines. It should be noted that, even though this work is here illustrated for the analysis of subcortical brain changes in neurological diseases, it can be extended to general multimodal multivariate analysis, such as to imaging-genetics studies.

The framework is benchmarked on synthetic data (section 3.1). It is then applied to the analysis of subcortical thickness and shape features across diseases from multi-centric, multi-database data including: Alzheimer's disease (AD), progressive and non-progressive mild cognitive impairment (MCIc, MCInc), Parkinson's disease (PD) and healthy individuals (HC) (section 3.2).

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Also, from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit `www.ppmi-info.org`. This research has been conducted using the UK Biobank Resource.

2 Methods

Biomedical data is assumed to be partitioned across different centers restricting the access to individual information. However, centers can individually share model parameters and run pipelines for feature extraction.

We denote the *global* data (e.g., image arrays) and covariates (e.g., age, sex information) as respectively $\mathbf{X}$ and $\mathbf{Y}$, obtained by concatenating respectively data and covariates of each center. Although these data matrices cannot be computed in practice, this notation will be used to illustrate the proposed methodology. In the global setting, variability analysis can be performed by analyzing the *global data covariance matrix* $\mathbf{S}$.

For each center $c \in \{1, \dots, C\}$ with N_c subjects each, we denote by $\mathbf{X}_c = (\mathbf{x}_i)_{i=1}^{N_c}$ and $\mathbf{Y}_c = (\mathbf{y}_i)_{i=1}^{N_c}$ the *local* data and covariates. The feature-wise mean and standard deviation vectors of each center are denoted as $\bar{\mathbf{x}}_c$ and σ_c.

The proposed framework is illustrated in Figure 1 and discussed in section 2.1. It is based on three main steps: 1) data standardization, 2) correction from confounding factors and 3) variability analysis.

Data standardization is a data pre-processing step, aiming to enhance the stability of the analysis and easing the comparison across features. In practice, each feature is mapped to the same space by centering data feature-wise to zero-mean and by scaling to unit standard deviation. However, this is ideally performed with respect to the statistics from the whole study (*global statistics*). This issue is addressed by proposing a distributed standardization method in section 2.1.1.

Confounding factors have a *biasing* effect on the data. To correct for this bias, it is usually assumed a linear effect of the confounders $\widehat{\mathbf{X}} = \mathbf{YW}$, that must be estimated and removed. However, for a distributed scenario, computing $\mathbf{W}$ is not straightforward, since the global data matrix cannot be computed. We propose in section 2.1.2 to use *Alternating Direction Method of Multipliers* (ADMM) to estimate a matrix $\widetilde{\mathbf{W}}$ shared among centers, closely approximating $\mathbf{W}$. In particular, we show that $\widetilde{\mathbf{W}}$ can be estimated in a federated way, without sharing local data $\mathbf{X}_c$ nor covariates $\mathbf{Y}_c$.

Finally, through federated principal component analysis (fPCA), we obtain a low dimensional representation of the full data without ever sharing any center's individual information $\mathbf{X}_c, \mathbf{Y}_c$ (section 2.1.3).

2.1 Federated Analysis Framework

2.1.1 Standardization

The mean and standard deviation vectors can be initialized to $\bar{\mathbf{x}}_0 = 0$ and $\bar{\sigma}_0 = 0$. They can be iteratively updated with the information of each center by following standard forms [8], by simply transmitting the quantities $\bar{\mathbf{x}}_c$ and σ_c from center to center. For each center the scaled data is denoted as $\widehat{\mathbf{X}}_c$ and keeps the dimensions of $\mathbf{X}_c$.

2.1.2 Correction from confounding factors

Under the assumption of a linear relationship between data and confounders, the parameters matrix $\mathbf{W}$ can be estimated via *ordinary least squares*, through the minimization of the error function $f(\mathbf{W}) = \left\| \mathbf{Y} - \widehat{\mathbf{X}}\mathbf{W} \right\|^2$.

In a distributed setting, this approach can be performed locally in each center, ultimately leading to C independent solutions. However, this would introduce a bias in the correction, as covariates are accounted for differently across centers.

To solve this issue, we propose to constrain the local solutions to a global one shared across centers. In this way, the subsequent correction can be consistently performed with respect to the estimated global parameters. Thus, we can formulate the problem of constrained regression via ADMM [9].

For a given error function $f_c(\mathbf{W}_c) = \left\|\mathbf{Y}_c - \widehat{\mathbf{X}}_c\mathbf{W}_c\right\|^2$ associated with each center c and constrained to a estimated global matrix of weights $\widetilde{\mathbf{W}}$ we can pose:

$$\text{minimize} \sum_{c=1}^{C} f_c(\mathbf{W}_c), \quad \text{subject to}\, \mathbf{W}_c = \widetilde{\mathbf{W}}, \quad \forall c.$$

As this is a constrained minimization problem, the extended Lagrangian can be calculated as a combination of the parameters from each center (eqn. 1).

$$L_\rho(\mathbf{W}, \widetilde{\mathbf{W}}, \alpha) \quad = \quad \sum_{c=1}^{C}\Big(f_c(\mathbf{W}_c) \quad + \quad \langle \alpha_c, \mathbf{W}_c \quad - \quad \widetilde{\mathbf{W}}\rangle \quad + \quad \frac{\rho}{2}\left\|\mathbf{W}_c - \widetilde{\mathbf{W}}\right\|_2^2\Big) \quad (1)$$

Where ρ is a penalty factor (or *dual update step length*) regulating the minimization step length for $\mathbf{W}$ and $\widetilde{\mathbf{W}}$. α is a *dual variable* to decouple the optimization of $\mathbf{W}$ and $\widetilde{\mathbf{W}}$.

Optimization is performed as follows: i) Each center independently calculates the local parameters $\mathbf{W}_c$ and α_c (eqn. 2 and 3); ii) the parameters $\mathbf{W}_c$ and α_c are shared to estimate the global parameters $\widetilde{\mathbf{W}}$ (eqn. 4). We note that this last step is performed without sharing either local data or covariates. The parameters $\widetilde{\mathbf{W}}$ are subsequently re-transmitted to the centers and the whole procedure is iterated until convergence:

$$\mathbf{W}_c^{(k+1)} := \arg\min_{\mathbf{W}_c} L_\rho(\mathbf{W}_c, \widetilde{\mathbf{W}}^{(k)}, \alpha_c^{(k)}) = \Big(\widehat{\mathbf{X}}_c'\widehat{\mathbf{X}}_c + \frac{\rho}{2}\mathbf{I}\Big)^{-1}\Big(\widehat{\mathbf{X}}_c'\mathbf{Y}_c - \frac{1}{2}\alpha_c^{(k)} + \frac{\rho}{2}\tilde{\mathbf{W}}_c^{(k)}\Big) \quad (2)$$

$$\alpha_c^{(k+1)} := \alpha_c^{(k)} + \rho\big(\mathbf{W}_c^{(k+1)} - \widetilde{\mathbf{W}}^{(k+1)}\big) \quad (3)$$

$$\widetilde{\mathbf{W}}^{(k+1)} \quad := \quad \arg\min_{\widetilde{\mathbf{W}}} L_\rho(\mathbf{W}_c^{(k+1)}, \widetilde{\mathbf{W}}, \alpha_c^{(k)}) \quad = \quad \frac{1}{C}\sum_{c}^{C}\left(\frac{\alpha_c^{(k)}}{\rho} \quad + \quad \mathbf{W}_c^{(k+1)}\right) \quad (4)$$

After convergence, $\widetilde{\mathbf{W}}$ is shared across centers, and used to consistently account for covariates by subtracting their effect from the structural data to obtain the corrected observation matrix: $\mathbf{E}_c = \widehat{\mathbf{X}}_c - \mathbf{Y}_c\widetilde{\mathbf{W}}$.

2.1.3 Federated PCA (fPCA)

Principal components analysis (PCA) is a standard approach for dimensionality reduction assuming that the largest amount of information is contained in the directions $\mathbf{U}$ (components) of greater

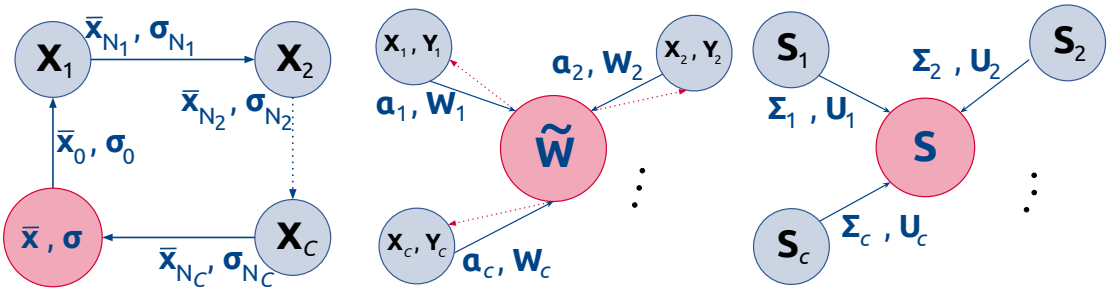


Figure 1: Data flow to obtain: (a) the global statistics $\bar{\mathbf{x}}$ and σ, (b) the shared parameter matrix $\widehat{\mathbf{W}}$ to correct from covariates and (c) the approximated global covariance matrix $\mathbf{S}$. Red node: master; blue nodes: local centers. Arrows denote the data flows from centers (blue) and from the master (red).

Database (total)	**ADNI (802)**				**MIRIAD (68)**		**PPMI (232)**	**UK Biobank (208)**
Group	HC	MCInc	MCIc	AD	HC	AD	PD	HC
N (females)	109 (115)	62 (119)	78 (130)	89 (100)	11 (12)	26 (19)	85 (147)	116 (92)
Age ± sd	75.79 (4.99)	74.93 (7.72)	74.54 (7.09)	75.19 (7.48)	69 (7.18)	69.17 (7.06)	60.69 (8.95)	60.72 (7.52)

Table 1: Data used in this study. Each study here represents an independent center. The centers are jointly analyzed through the federated analysis proposed in Section 2.1.

variability. Data can be thus represented by projecting on the low-dimensional space spanned by the main components: $\widehat{\mathbf{E}} = \mathbf{EU}$.

From the eigen-value decomposition of the global covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}'$, the first m-eigen-modes $\mathbf{U} = (\mathbf{u}_j)_{j=1}^m$ provide a low-dimensional representation of the overall variation in $\mathbf{E}$. In our federated setting, we note that $\mathbf{S}$ is the algebraic sum of the *local covariance matrices* $\mathbf{S} = \mathbf{E}\mathbf{E}' = \sum_{c=1}^{S} \mathbf{E}_c\mathbf{E}_c'$. Based on this observation, Lorenzi *et* al. proposed to share only the eigen-modes and values of the covariance matrix of each center avoiding the access to individual data [6]. However, sharing the local-covariance-matrices can still be prohibitive as the dimension is $(N_{\text{features}} \times N_{\text{features}})$. For this reason, it was proposed to further reduce the dimensionality of the problem by sharing only the principal eigen-components associated with the local covariance matrices: $\mathbf{S} \approx \sum_{c=1}^{C} \mathbf{U}_c\mathbf{\Sigma}_c^2\mathbf{U}_c'$. From the practical point of view, computing the eigen-components can be efficiently performed by solving the eigen-problem associated with the matrix $(\mathbf{X}_c\mathbf{X}_c')^2$ which is usually of much smaller dimension $(N_c \times N_c)$ [10].

In what follows, the number of components shared across centers is automatically defined by fixing a threshold of 80% on the associated *explained variability* contained in $\mathbf{\Sigma}_c$.

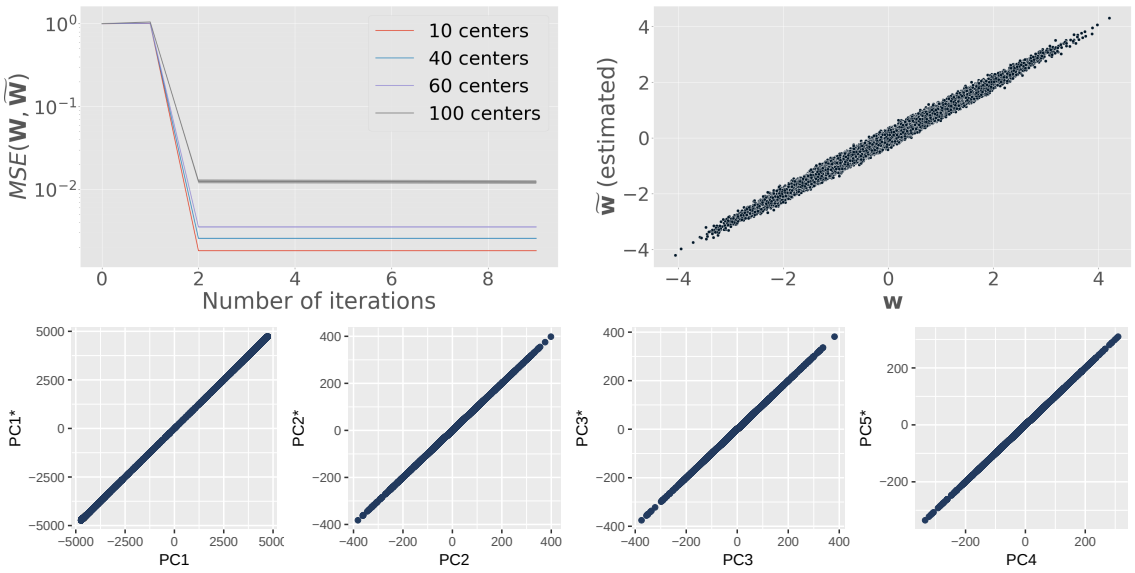


Figure 2: **Top-left:** Mean square error (MSE) between $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ for different numbers of centers. $N = 2400$, $N_{\text{features}} = 50,000$ and $\dim(\mathbf{y}) = 20$. **Top-right:** Single-column of $\mathbf{W}$ vs $\widetilde{\mathbf{W}}$ for $C = 100$. **Bottom:** Principal components (PC) vs federated ones (PC*) for 100 centers.

3 Experiments

3.1 Synthetic Data

We randomly generated $\mathbf{Y}$ and $\mathbf{W}$ matrices. Data matrix was subsequently computed as $\mathbf{X} = \mathbf{YW}$, and corrupted with Gaussian noise $\mathcal{N}(0, \sigma)$, with σ set to 20% of $\|\mathbf{Y}\|$. Then, $\mathbf{X}$ and $\mathbf{Y}$ were split in C centers of equal sample size. Our federated framework was then applied for each scenario across 200 folds, and convergence analyzed as shown in Figure 2.

3.2 Real Data: Neuroimaging

Data. T1-weighted MRI scans at baseline were analyzed from several research databases (table 1). In total, we included data for 455 controls (HC), 181 with non-progressive MCI (MCInc), 208 progressive (MCIc), 234 Alzheimer's disease (AD), 232 with Parkinson's disease (PD).

Feature extraction. ENIGMA Shape Analysis was applied to the MRI data of each center [11, 12]. In our analysis we extracted: a) radial distance (an approximate measure of thickness) and, b) the log of the Jacobian determinant (surface area dilation/contraction) for each vertex of the following subcortical regions: hippocampi, amygdalae, thalami, pallidum, caudate nuclei, putamen and accumbens nuclei. The overall data dimension is of 54,240 features.

Federated analysis. Each database of table 1 was modeled as an independent center. *Sex*, *Age* and Age^2 were used to correct the vertex-wise shape data according to 2.1.1 and 2.1.2. For ADMM, convergence was ensured through 10 iterations. Finally, the analysis of the variability was

performed according to 2.1.3.

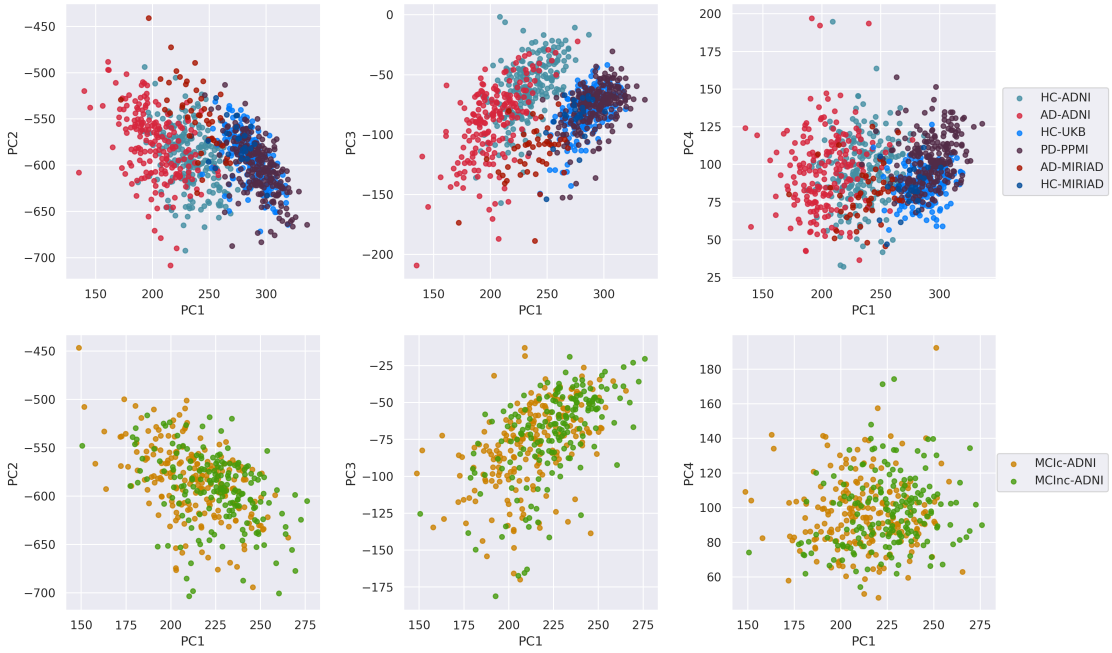


Figure 3: Data projected on the first 4 components. AD vs controls from different centers (top). MCI progressive and stable from ADNI (bottom). Federated PCA was performed on the whole data obtained from the 4 centers (table 1).

Results. The projection in the latent space spanned by the federated principal components is shown in Figure 3. To ease visualization, the projection for MCI converters and those who remained stable is shown in the bottom panel. Figure 4 shows the weight maps associated to the first principal component. We note that principal components 1 to 3 identify a variability from healthy to AD consistent across centers. Moreover, healthy ADNI participants are in between the AD subjects and the rest of the population. This result may denote some residual effect of *Age* on the resulting imaging features, even after correction. Interestingly, the issue of 'leaking' spurious variability of confounders after correction has been already reported in a number of multi-centric studies, and is matter of ongoing research [13, 14]. Finally we note that PD subjects are generally similar to the healthy individuals with respect to the modelled subcortical information.

4 Conclusions

In this work we proposed, tested, and validated a fully consistent framework for federated analysis of distributed biomedical data. Further developments of this study will extend the proposed analysis to large-scale imaging genetics data, such as in the context of the ENIGMA meta-study.

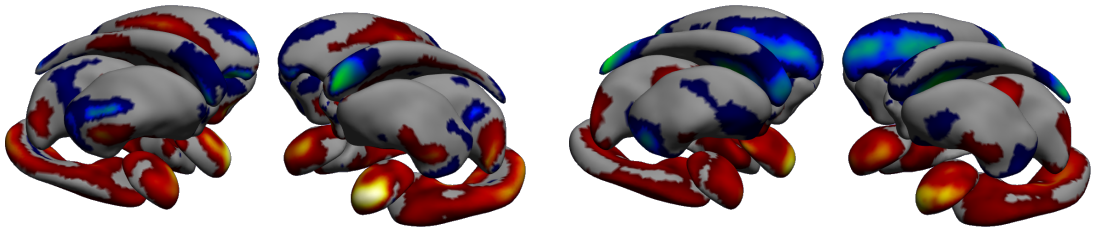


Figure 4: First principal component estimated with the proposed federated framework. The component maps prevalently hippocampi and amigdalae. **Left:** Thickness. **Right:** Log-Jacobians.

5 Acknowledgments

This work was supported by the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01 (project Meta- ImaGen). A.A. holds an Medical Research Council eMedLab Medical Bioinformatics Career Development Fellowship. This work was supported by the Medical Research Council [grant number MR/L016311/1]. Additional support is from NIH grants RF1AG051710, R01MH116147 and R56AG058854. The full list of acknowledgments for the data providers is available at `https://hal.inria.fr/hal-01895800/document`.

References

[1] Junfeng Sun et al., "Meta-analysis of Clinical Trials," *Principles and Practice of Clinical Research*, pp. 317–327, jan 2018.

[2] Paul M. Thompson et al., "The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data," *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 153–182, 2014.

[3] Bradley T Baker, Rogers F Silva, Vince D Calhoun, Anand D Sarwate, and Sergey M Plis, "Large scale collaboration with autonomy: Decentralized data ica," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015.

[4] Jing Ming, Eric Verner, Anand Sarwate, Ross Kelly, Cory Reed, Torran Kahleck, Rogers Silva, Sandeep Panta, Jessica Turner, Sergey Plis, et al., "Coinstac: Decentralizing the future of brain imaging analysis," *F1000Research*, vol. 6, 2017.

[5] Sergey M. Plis, Anand D. Sarwate, Dylan Wood, Christopher Dieringer, Drew Landis, Cory Reed, Sandeep R. Panta, Jessica A. Turner, Jody M. Shoemaker, Kim W. Carter, Paul Thompson, Kent Hutchison, and Vince D. Calhoun, "Coinstac: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data," *Frontiers in Neuroscience*, vol. 10, pp. 365, 2016.

[6] Marco Lorenzi et al., "Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study," p. 1016016, 1 2017.

[7] Marco Lorenzi et al., "Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease, evidence from functional prioritization in imaging genetics," *Proceedings of the National Academy of Sciences*, vol. 115, no. 12, pp. 3162–3167, 3 2018.

[8] B. P. Welford, "Note on a Method for Calculating Corrected Sums of Squares and Products," *Technometrics*, vol. 4, no. 3, pp. 419, 8 1962.

[9] Stephen Boyd, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[10] Keith J. Worsley et al., "Comparing functional connectivity via thresholding correlations and singular value decomposition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 913–920, 2005.

[11] Benjamin S.C. Wade et al., "Mapping abnormal subcortical brain morphometry in an elderly HIV + cohort," *NeuroImage: Clinical*, vol. 9, pp. 564–573, 2015.

[12] Gennady V. Roshchupkin et al., "Heritability of the shape of subcortical brain structures in the general population," *Nature Communications*, vol. 7, pp. 1–8, 2016.

[13] Jacob Westfall and Tal Yarkoni, "Statistically controlling for confounding constructs is harder than you think," *PloS one*, vol. 11, no. 3, pp. e0152719, 2016.

[14] Stephen M Smith and Thomas E Nichols, "Statistical challenges in "Big Data" human neuroimaging," *Neuron*, vol. 97, no. 2, pp. 263–268, 2018.

Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study

Marco Lorenzi[1], Boris Gutman[2], Paul M. Thompson[2], Daniel C. Alexander[3], Sebasetien Ourselin[1] and Andre Altmann[1]

1.Translational Imaging Group, CMIC, University College London, London, UK

2. Imaging Genetics Center, University of Southern California, Marina del Rey, CA, USA

3.Centre for Medical Image Computing, University College London, London, UK

Originally published in:
12th International Symposium on Medical Information Processing and Analysis, 1016016, International Society for Optics and Photonics, 2017.

Abstract

State-of-the-art data analysis methods in genetics and related fields have advanced beyond massively univariate analyses. However, these methods suffer from the limited amount of data available at a single research site. Recent large-scale multi-centric imaging-genetic studies, such as ENIGMA, have to rely on meta-analysis of mass univariate models to achieve critical sample sizes for uncovering statistically significant associations. Indeed, model parameters, but not data, can be securely and anonymously shared between partners. We propose here partial least squares (PLS) as a multivariate imaging-genetics model in *meta-studies*. In particular, we propose an online estimation approach to partial least squares for the sequential estimation of the model parameters in data batches, based on an approximation of the singular value decomposition (SVD) of partitioned covariance matrices.We applied the proposed approach to the challenging problem of modeling the association between 1,167,117 genetic markers (SNPs, single nucleotide polymorphisms) and the brain cortical and sub-cortical atrophy (354,804 anatomical surface features) in a cohort of 639 individuals from the Alzheimer's Disease Neuroimaging Initiative. We compared two different modeling strategies (*sequential-* and *meta-PLS*) to the classic non-distributed PLS. Both strategies exhibited only minimal approximation errors of model parameters. The proposed approaches pave the way to the application of multivariate models in large scale imaging-genetics meta-studies, and may lead to novel understandings of the complex brain phenotype-genotype interactions.

1 Introduction

State-of-the-art data analysis methods in genetics and related fields have advanced beyond massively univariate analyses. However, these methods suffer from the limited amount of data available at a single research site. The reliability of multivariate models in imaging-genetics is usually hampered by the low sample size of the studies on the order of 100s of individuals, relatively to the large number of parameters, which is at least one order of magnitude higher.

Genetic variants often have only a small effect on disease risk or other quantitative phenotype such as measures derived from brain imaging. Thus, in order to gather sufficient statistical power to detect significant associations large samples sizes are required typically in the range of 10,000s subjects. Achieving this critical sample size is often hampered by practical considerations such as the need to transfer large volumes of data to one single research site and the bureaucratic burden associated with data transfer agreements. To circumvent data transfer large genetics and imaging-genetics consortia have relied on the concept of meta analysis. For instance, ENIGMA [1], which studies the association between brain phenotype and genotype on very large cohorts, relies on meta-analysis of mass univariate models to achieve critical sample sizes. In a meta-analysis only the results of a statistical test (i.e., p-value, effect size, standard error, sample size) are shared, but not the individual-level data. In this way, model parameters, but not data, can be securely and anonymously shared between partners. Indeed, meta-analysis represents the standard paradigm of modern large-scale clinical research projects, involving research and medical institutions with different data sharing policies and restrictions. Thus far, meta-analysis has been applied only to classic univariate associate tests. Hence, the development of powerful multivariate modeling approaches within a meta-analysis context is an impelling need to better model the complex brain phenotype-genotype interactions in very large cohorts. We propose here partial least squares (PLS) as a multivariate imaging-genetics model in meta-studies. In particular, inspired by early works on recursive partial least squares[2] we propose a novel iterative approach to PLS for the sequential estimation of the model parameters in data batches, based on the approximation of the singular value decomposition (SVD) of partitioned covariance matrices.

We propose two different meta-modeling strategies for parameter estimation and cross-validation, which are compliant with the anonymity restrictions: 1) the PLS model parameters are transmitted and updated by each centre sequentially (*sequential-PLS*), and 2) the PLS model parameters are independently estimated by each centre and subsequently merged (*meta-PLS*). While both strategies are asymptotically equivalent, their degree of approximation depends on the batch sample size, and on the number of latent components. We applied the proposed strategies to the challenging problem of modeling the multivariate association between 1,167,117 genetic markers (SNPs; single nucleotide polymorphisms) and the brain cortical and sub-cortical atrophy (354,804 anatomical surface features) in a cohort of 639 individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We compared *sequential-* and *meta-PLS* to the classic non-distributed PLS, assessing agreement in model parameters. Both strategies exhibited only minimal approximation errors.

The proposed approaches pave the way to the application of multivariate models in large scale imaging-genetics meta-studies, and may lead to novel understandings of the complex brain phenotype-genotype interactions.

2 PLS for the analysis of multimodal features in medical imaging

This section introduces notations and the theoretical context of PLS. Let $\mathbf{X} = \{x_i\}_1^{N_{sub}}$ and $\mathbf{Y} = \{y_i\}_1^{N_{sub}}$ be $n\_subjects \times n\_features$ observation matrices of features $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^m$ for N_{sub} individuals.

PLS is a standard approach for modeling the joint variation between $\mathbf{X}$ and $\mathbf{Y}$, and is classically formulated through the decomposition of the matrices $\mathbf{X}$ and $\mathbf{Y}$ by means of linear mappings $\mathbf{u}$

and $\mathbf{v}$. The mappings are optimised in order to maximise the covariance between the projections, $\mathbf{Xu}$, and $\mathbf{Yv}$.

Among the several versions of PLS [3, 4, 5, 6, 7], we focus here on the symmetric formulation of PLS computed through the SVD of the cross-covariance matrix $\mathbf{XY}' = \mathbf{UWV}'$. This approach has been inspired by the analysis of principal modes of variability in neuroimaging data [8] and has been popularized in the field of neuroimaging in the seminal works [5, 6], for the study of positron emission tomography (PET) and functional magnetic resonance images (fMRI) through the analysis of the associated eigen-modes of intensity variation.

The first k eigen-modes $\mathbf{U}' = \{\mathbf{u}_i\}_{i=1}^k$ and $\mathbf{V}' = \{\mathbf{v}_i\}_{i=1}^k$ provide a low-dimensional representation of the main correlation modes between $\mathbf{X}$ and $\mathbf{Y}$, where the relative components weights are informative of the relevance of specific features in explaining the overall variation. In spite of the apparently prohibitive computational cost of the SVD of the large covariance matrix $\mathbf{XY}'$ ($n\_features_X \times n\_features_Y$), Worsley and colleagues showed that the main eigen-modes of variation can be derived from the solution of the eigen-problem associated to the usually smaller ($n\_subjects \times n\_subjects$) matrix $\mathbf{XX}'\mathbf{YY}'$ [9]. Thanks to this contribution it has been shown that PLS can be successfully employed in the modeling of high-dimensional functional and structural brain connectivity [9], as well as of the joint variation between brain structure and function [10], and between imaging and genetic data [11, 12].

3 SVD-PLS in large-scale multi-centric studies

In this section we propose a novel approach to PLS within an online-learning framework. We assume that the set of observations is partitioned in clusters C_l, $l = 1, \ldots, N_l$, representing for instance different clinical centres, and we denote with $\mathbf{X}_c$ and $\mathbf{Y}_c$ the set of observations belonging to the cluster c.

We start by noting that, with reference to Figure 1, the global cross-covariance matrix $C = \mathbf{XY}'$ can be decomposed as the sum of cluster-specific covariances: $\mathbf{XY}' = \sum_l \mathbf{X}_l\mathbf{Y}_l'$. Therefore, in a meta-analysis context, the estimation of the SVD-PLS model does not require the access to individual data, and can be done by sharing the cross-covariance matrices C_l. However this operation may be still prohibitive, since the matrices C_l are as well of dimension $n\_features_X \times n\_features_Y$, therefore usually very large. However, this limitation can be overcome by replacing the matrices $\mathbf{X}_l\mathbf{Y}_l'$ by their approximation via SVD decomposition: $\mathbf{X}_l\mathbf{Y}_l' = \mathbf{U}_l\mathbf{W}_l\mathbf{V}_l'$, and thus by sharing only the first k_l eigen-vectors $\mathbf{u}_i^{(l)}$ and $\mathbf{v}_i^{(l)}$, along with the respective eigen-values $\{w_i^{(l)}\}$. The subsequent analysis can then be carried on the approximated covariance $\tilde{C} = \sum_l \mathbf{U}_l\mathbf{W}_l\mathbf{V}_l' \sim \sum_l \mathbf{X}_l\mathbf{Y}_l'$. It is worth noting that the SVD of the matrix $\tilde{C}$ does not require the explicit computation of the products $\mathbf{U}_l\mathbf{W}_l\mathbf{V}_l'$, which would be computationally cumbersome. Indeed, it is straightforward to show that $\tilde{C} = \tilde{\mathbf{X}}\tilde{\mathbf{Y}}'$, where the columns of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are respectively the components $\{\mathbf{u}_i^{(l)}\}$, and $\{w_i^{(l)}\mathbf{v}_i^{(l)}\}$, $i = 1, \ldots, k_l$, and $l = 1, \ldots, N_l$. In this way the SVD of $\tilde{C}$ can still be efficiently computed with the approach proposed by Worsley and colleagues [9].

This numerical scheme motivates the definition of the follwing online-learning approaches to PLS, based on two different learning strategies, and denoted respectively by *sequential* and *meta*-PLS.

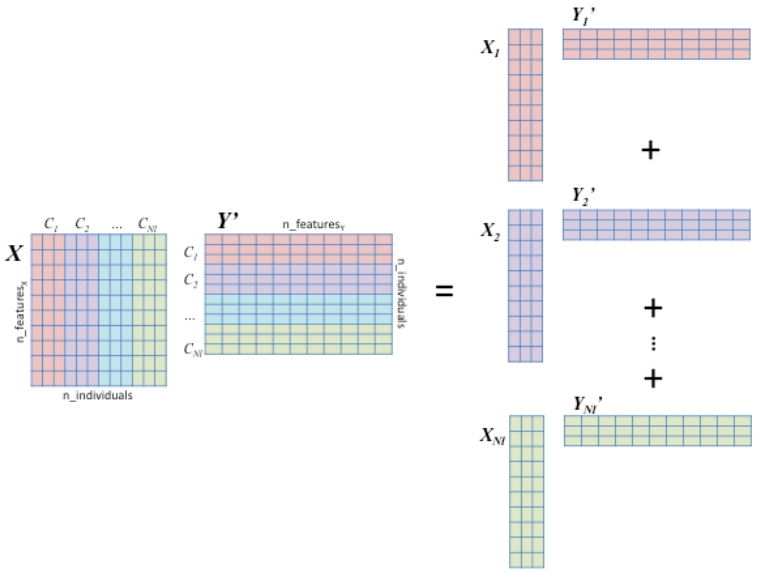


Figure 1: The cross-covariance matrix of partitioned data can be decomposed as the sum of cluster-specific covariances.

3.1 Sequential-PLS

In sequential-PLS we start from an initial approximation of the first k_0 eigen-components of the cross-covariance matrix: $\{\mathbf{u}_i^{(0)}\}, \{\mathbf{v}_i^{(0)}\}$, and $\{w_i^{(0)}\}$, $i = 1, \ldots, k_0$. The components are then transmitted to the centre C_1, and are used to generate the augmented matrices $\tilde{\mathbf{X}}_1 = [\mathbf{X_1}, \{\mathbf{u}_i^{(0)}\}]$, and $\tilde{\mathbf{Y}}_1 = [\mathbf{Y_1}, \{w_i^{(0)}\mathbf{v}_i^{(0)}\}]$. The SVD of the cross-covariance matrix $\tilde{C}_1 = \tilde{\mathbf{X}_1}\tilde{\mathbf{Y}_1}'$ thus estimates the updated components $\{\mathbf{u}_i^{(1)}\}, \{\mathbf{v}_i^{(1)}\}$, and $\{w_i^{(1)}\}$, $i = 1, \ldots, k_1$, that are subsequently transmitted to the next centres and updated in an iterative fashion.

This strategy allows at each step to estimate the model parameters by exploiting the data at each centre. The approximation of sequential-PLS arises from the degree of approximation of the transmitted components $\mathbf{u}^{(l)}$, $\mathbf{v}^{(l)}$, and $w^{(l)}$ in the factorization of the covariance $\tilde{C}_l$, that can be eventually negligible by sharing an adequate number of SVD components.

The drawback of sequential-PLS lies in the necessity of re-estimating the whole model in case of changes occurring at a single centre. A second practical drawback is that the centers have to coordinate themselves or have to be coordinated by a "moderator".

3.2 Meta-PLS

In meta-PLS each centre C_l independently estimates the eigen-components $\{\mathbf{u}_i^{(l)}\}, \{\mathbf{v}_i^{(l)}\}$, and $\{w_i^{(l)}\}$, $i = 1, \ldots, k_l$ approximating the cross-covariances $\mathbf{X}_l\mathbf{Y}_l'$. The different components are then concatenated to form matrices $\tilde{\mathbf{X}} = [\{\mathbf{u}^{(l)}\}]$ and $\tilde{\mathbf{Y}} = [\{w^{(l)}\mathbf{v}^{(l)}\}]$, and the eigen-components are finally obtained from the SVD of the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{Y}}'$.

Similarly as in sequential-PLS, the approximation introduced in meta-PLS depends on the

	Healthy	AD
# individuals	401	238
Sex (% females)	49	45
MMSE	29.1	23.2
Education (years)	16.36	15.34
APOE4 (% 0,1,2)	72, 26, 2	31, 48, 21

Table 1: Summary socio-demographic, clinical and genetic information. MMSE: mini-mental state examination.

approximations made at each centre. However, since the full model is estimated on the joint components in $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, it does not directly exploit the data at each centre.

An advantage of meta-PLS is that it can easily accommodate for eventual changes occurring in a single centre, which can eventually re-transmit the data without affecting the model estimation in the others sites.

4 Numerical approximation

The quality of the approximation of the proposed strategies clearly depends on the number of eigen-components chosen for approximating the cross-covariance matrix at each site. In what follows the number of eigen-components is automatically estimated from the set of singular values, in order explain the 90% of the overall variability of the entire data available at each centre.

5 Model validation in a controlled setting: an imaging genetics case study

We tested the proposed sequential- and meta-PLS strategies in the context of modeling the joint variability in imaging-genetics, in the same application proposed in [12].

5.1 Data processing

We selected genotype and phenotype data available in the ADNI-1/GO/2 datasets for 639 subjects. At time of imaging/study entry subjects were diagnosed as healthy individuals (N=401), and Alzheimer patients (N=238). Summary socio-demographic, clinical and genetic information are available in Table 1.

The imaging phenotype consisted of the individuals' baseline brain cortical thickness maps estimated by Freesurfer [13], and the bilateral radial thickness maps for hippocampi and amygdalae [14]. The imaging component comprises 327,684 cortical and 27,120 subcortical features per subject. These raw thickness values were normalised by covarying for age, total intracranial volume, and sex. Subsequently, data were standardised by group-wise mean and standard deviation computed in the pooled group of healthy and AD individuals.

Genotype data (Illumina Human610-Quad BeadChip for ADNI-1, and Illumina Human Omni Express for ADNI-2/GO) was downloaded from the ADNI website and preprocessed with PLINK [15]. Standard quality control (QC) parameters were used to filter SNPs: Minor Allele Frequency

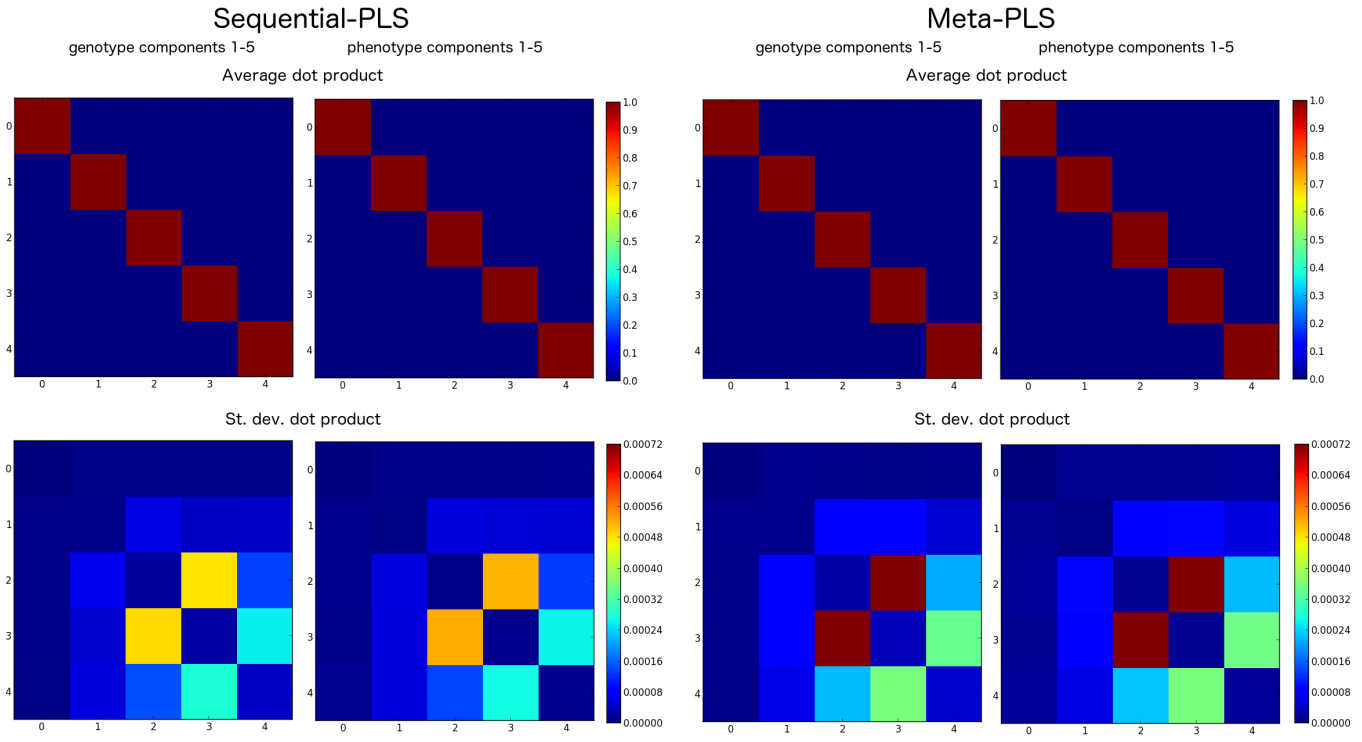


Figure 2: Average (top) and standard deviation (bottom) for the dot product between the components estimated with sequential- and meta-PLS, and the components estimated with the classic non-distributed PLS. Both strategies lead to minimal deviation from the benchmark.

(MAF) < 0.01, Genotype Call Rate <95% and Hardy-Weinberg Equilibrium $< 1 \times 10^{-6}$. Finally, SNPs passing QC were imputed to the HapMap III reference panel and further quality controlled to keep only high quality imputed SNPs (i.e., MAF > 0.01 and imputation quality score > 0.3). Missing individual SNPs were replaced by the group-wise median. The genotype features consisted in the individuals' minor allele counts for each of the resulting 1,167,126 SNPs in chromosomes 1 to 22. The resulting allele counts were finally standardised by group-wise mean and standard deviation computed in the pooled group of healthy and AD individuals.

5.2 Statistical analysis

The data was randomly partitioned in two non-overlapping groups (each of size n= 319) in order to simulate independent centres. Within this simulated setting, sequential- and meta-PLS were applied to estimate the respective model parameters. The results were compared to those obtained with the classic non-distributed PLS, in terms of dot-product between eigen-components (which quantifies the angle between the spanned eigen-spaces), and of the absolute feature-wise error between the components weights, measured as $\Sigma_i \frac{|w_i|-|\tilde{w}_i|}{|w_i|}$, where w_i and $\tilde{w}_i$ are features for

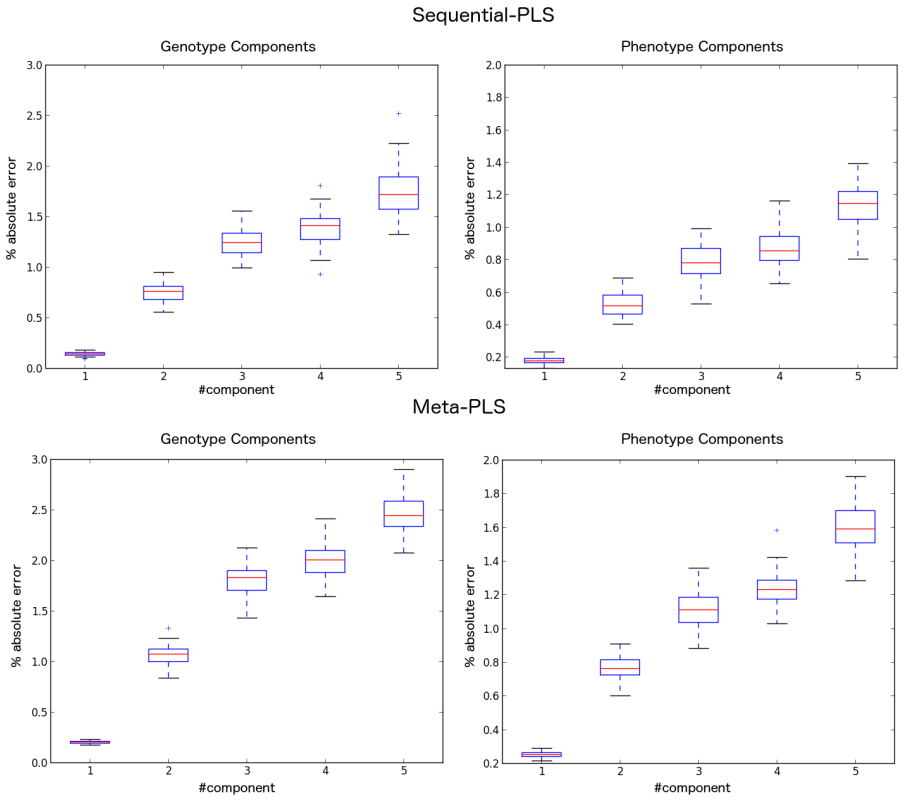


Figure 3: Absolute feature-wise error between the components weights estimated with sequential- and meta-PLS and with the classic non-distributed PLS. Both strategies lead to minimal differences with respect to the benchmark.

respectively non-distributed and online PLS schemes. The assessment was performed on the first 5 principal eigen-components, and the whole procedure was repeated 50 times with varying data partitions.

6 Results: component stability

Figure 2 shows the component-wise dot-product between the proposed strategies and the non-distributed PLS, averaged across folds. For all the considered cases the product matrix is diagonal, to indicate that both sequential- and meta-PLS lead to negligible deviations from the benchmark. Nevertheless, the variability introduced by meta-PLS is slightly higher. The approximation quality of the proposed online-learning schemes is confirmed by the feature-wise absolute error between components weights shown in the boxplots of Figure 3. Errors are generally of small magnitude, and increase for the higher components. In particular, sequential-PLS generally leads to slightly better approximation than meta-PLS.

On average 282 components were estimated at each centre in order explain 90% of the overall local variability.

7 Conclusions

In this work we explored an innovative approach to multivariate modeling in a meta-analysis context. We showed that classical SVD-PLS can be naturally extended to online-learning schemes by leveraging on simple algebraic properties of partitioned covariance matrices. We compared two different modeling strategies (*sequential-* and *meta-PLS*) to the classic non-distributed PLS. We show that the methods have great promise for our target application, imaging genetics. In a preliminary study of 639 subjects from the ADNI dataset with over 10^5 brain MRI-based imaging features and 10^6 genetic variants, we are able to demonstrate good convergence properties of both meta-PLS and sequential-PLS. Indeed, the approximation errors, as measured by overall PLS component compatibility, are negligible, while the individual feature weight error remains within 3%. This is a remarkable consistency in a dataset with thousands of times more features than subjects. The proposed approaches thus pave the way to the application of multivariate models in large scale imaging-genetics meta-studies, and may lead to novel understandings of the complex brain phenotype-genotype interactions. To date, there have been many successful Genome-Wide Association (GWAS), or mass-univariate studies, relying on meta-analysis. For a number of practical reasons, the latter has become the Modus Operandi of large genetics consortia, even beyond brain imaging. Yet, we are not aware of any multi-centre meta-analytic studies using multivariate techniques, such as PLS. Our approach has tremendous potential to lead to new discoveries of associations between brain imaging phenotypes and common genetic variants, particularly where multi-SNP and multi-phenotype interactions are at play.

The accuracy of the proposed schemes critically depends on the low-rank approximation at each centre. Extensions of this work will aim at investigating the relationship between the number of components shared by each centre and the overall model approximation. Another important point that will be tackled in future studies concerns the study of online cross-validation schemes for estimating confidence intervals for sequential- and meta- PLS parameters. This aspect is critical for the inference and interpretation of modeling results in imaging-genetics studies.

There is an analogy between the proposed approach and the recent group-PCA method proposed in fMRI analysis [16]. In both cases we aim at an approximation of the overall covariance matrix by serial updates of eigen-components estimated in data batches. In this work we extend this idea to the multimodal setting, and we develop the theory necessary for the implementation of the model in meta-analysis.

Finally, we hope this development will soon lead to real imaging-genetics discoveries. As part of future development, we plan to integrate meta-PLS and sequential-PLS into a large imaging genetics consortium study.

References

[1] Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., et al., "The ENIGMA consor-

tium: large-scale collaborative analyses of neuroimaging and genetic data," *Brain imaging and behavior* **8**(2), 153–182 (2014).

[2] Qin, S. J., "Recursive PLS algorithms for adaptive data modeling," *Computers & Chemical Engineering* **22**, 503–514 (1998).

[3] Wold, H., "Estimation of principal components and related models by iterative least squares. multivariate analysis. edited by: Krishnaiaah pr. 1966."

[4] Martens, H. and Naes, T., [*Multivariate calibration*], John Wiley & Sons (1992).

[5] McIntosh, A., Bookstein, F., Haxby, J. V., and Grady, C., "Spatial pattern analysis of functional brain images using partial least squares," *Neuroimage* **3**(3), 143–157 (1996).

[6] Worsley, K. J., "An overview and some new developments in the statistical analysis of PET and fMRI data," *Human Brain Mapping* **5**(4), 254–258 (1997).

[7] McIntosh, A. R. and Lobaugh, N. J., "Partial least squares analysis of neuroimaging data: applications and advances," *Neuroimage* **23**, S250–S263 (2004).

[8] Friston, K., Frith, C., Liddle, P., and Frackowiak, R., "Functional connectivity: the principal-component analysis of large (PET) data sets," *Journal of Cerebral Blood Flow & Metabolism* **13**(1), 5–14 (1993).

[9] Worsley, K. J., Chen, J.-I., Lerch, J., and Evans, A. C., "Comparing functional connectivity via thresholding correlations and singular value decomposition," *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **360**(1457), 913–920 (2005).

[10] Lorenzi, M., Simpson, I. J., Mendelson, A. F., Vos, S. B., Cardoso, M. J., Modat, M., Schott, J. M., and Ourselin, S., "Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations," *Scientific reports* **6**(22161) (2016).

[11] Lorenzi, M., Gutman, B., Hibar, D. P., Altmann, A., Jahanshad, N., Thompson, P. M., and Ourselin, S., "Partial least squares modelling for imaging-genetics in Alzheimer's disease: Plausibility and generalization," in [*2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*], 838–841, IEEE (2016).

[12] Lorenzi, M., Gutman, B., Altmann, A., Hibar, D. P., Jahanshad, N., Thompson, P. M., and Ourselin, S., "Linking gene pathways and brain atrophy in Alzheimer's disease." Alzheimer's Association International Conference (AAIC 2016) (2016).

[13] Dale, A. M., Fischl, B., and Sereno, M. I., "Cortical surface-based analysis: I. segmentation and surface reconstruction," *Neuroimage* **9**(2), 179–194 (1999).

[14] Gutman, B. A., Hua, X., Rajagopalan, P., Chou, Y.-Y., Wang, Y., Yanovsky, I., Toga, A. W., Jack, C. R., Weiner, M. W., Thompson, P. M., et al., "Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features," *Neuroimage* **70**, 386–401 (2013).

[15] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics* **81**(3), 559–575 (2007).

[16] Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., and Beckmann, C. F., "Group-PCA for very large fMRI datasets," *NeuroImage* **101**, 738–749 (2014).